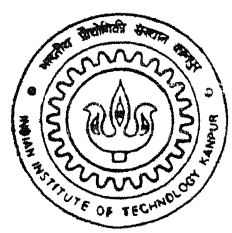
ESTIMATION OF FORMANT PARAMETER VARIATIONS USING LINEAR PREDICTION

by

ABHIRUP DAS BARMAN



DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KANPUR

April, 1999

ESTIMATION OF FORMANT PARAMETER VARIATIONS USING LINEAR PREDICTION

A Thesis Submitted

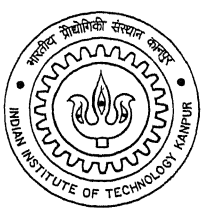
in Partial Fulfilment of the Requirements

for the Degree of

MASTER OF TECHNOLOGY

by

ABHIRUP DAS BARMAN



to the

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

April, 1999

CERTIFICATE

This is to certify that the work, "**Estimation of Formant Parameter Variations using Linear Prediction**", has been carried out by Abhirup Das Barman under our supervision and it has not been submitted elsewhere for a degree.

Preeti Rao

Dr. Preeti Rao
Asst. Professor
Dept. of Electrical Engg.
Indian Institute of Technology
Kanpur

G Sharma

Dr. Govind Sharma
Associate Professor
Dept. of Electrical Engg.
Indian Institute of Technology
Kanpur

Dedications

This thesis is dedicated to my family for their blessings and encouragement, which has helped me to complete the present work smoothly.

Acknowledgements

I express my gratitude to my thesis supervisor Dr. Preeti Rao for her active guidance at every stage of this thesis work. I would also like to thank my co-supervisor Dr. Govind Sharma, who offered many suggestions for the improvement of the thesis work. My special thanks go to Doordarshan Directorate and without their assistance it would not have been possible for me to complete this work.

I wish to thank all those with whom I had a nice time at IIT Kanpur.

Abstract

Estimation of vocal tract characteristics from the speech signal is an important problem. The properties of the vocal tract as represented by the speech signal formant parameters vary in time due to the movement of the articulators as well as due to the vocal fold oscillations in each pitch period. In this work the problem of the estimation of formant parameters as they vary in time due to the changing source-tract coupling during the glottal cycle is addressed. A method using covariance based linear predictive analysis is studied. Continuously varying formant trajectories are obtained by peak-picking of linear prediction spectrum from a sliding short data window. In order to interpret the frequency estimates correctly, the instants of significant excitation corresponding to glottal closure and opening are determined using the prediction error and the log determinant methods. Experimental results using simulated as well as natural speech data are presented. In several cases of natural speech vowels, clear increases of formant frequency are observed in the open phase compared with the closed phase.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The main task of signal processing systems are to generate, detect and interpret signals carrying valuable informations. One of the most powerful techniques is the method of linear predictive analysis. This method has become the predominant technique for estimating basic speech parameters, e.g., pitch, formants, spectra and representing speech for low data rate transmission. The importance of this method lies both in its ability to provide extremely accurate estimates of speech parameters and its relative speed of computation.

The acoustic theory of speech production leads to a variety of ways of representing the speech signal. When air from lungs flows through trachea at the constriction, pressure falls off according to Bernoulli's law. So opening between the vocal cords (glottis) comes together and completely constricts air flow (called closed phase). As a result pressure builds up behind the vocal cords and when it builds up sufficiently it forces vocal cords to open and thus allows air to flow through the glottis again (called open phase). This cycle is repeated. Thus vocal cords entered in a condition of sustained oscillation. The rate at which glottis opens and closes (called pitch frequency) are controlled by air pressure from lungs, the tension and

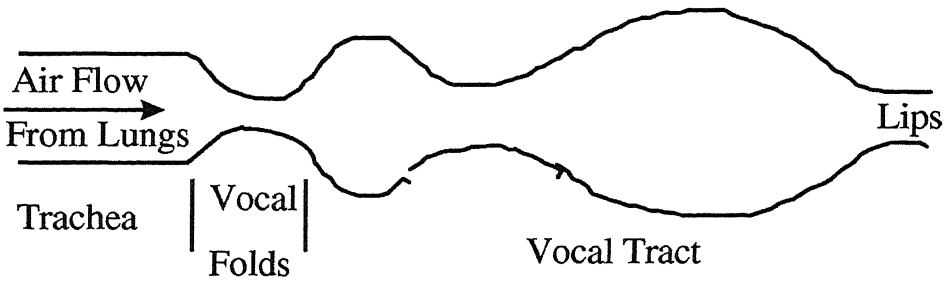


Figure 1.1: Schematic representation of the vocal system

stiffness of vocal cords and the area of the glottal opening. So glottal flow is automatically chopped into closed and open phases which can be thought of as two instants of excitation, one triggered as glottal closes and another of relatively low strength triggered as glottal opens. When the vocal folds are closed, output air flow is a freely decaying oscillation through the vocal tract.

The conventional linear predictive analysis models voiced speech signal as the output of an all pole system to a quasi periodic train of impulses, the period equal to the pitch period. This conventional model assumes that there is a single instant of excitation in a pitch period. This assumption is made because the major excitation of the vocal tract occurs at glottal closure. This is a strong spike-like excitation just preceding the glottal closure. This explains why this is called an instant of significant excitation. Hence this model does not take into account the secondary excitation that occurs when glottal is open. Significant excitation, although less in magnitude, persists inside glottal open and couples with secondary excitation. So it is reasonable that each pitch period be represented by two steady states one after glottal closure and one after glottal open separated by short intervals (transitions)

during which changes takes place at the glottis. The properties of the vocal tract changes even within a pitch period because of opening and closing of glottis.

Characterisation of Vocal Tract System

The shape of the vocal tract system is determined by the positions of the articulators (lip, tongue, jaw etc.). The vocal-tract shape is difficult to derive from the speech signal. Hence the free resonance of the vocal tract system called formant is used to characterise the vocal tract system. Broadly a formant is described by the three parameters:

1. the formant frequency
2. the formant bandwidth
3. the formant amplitude

Formant Variations

Formant can vary in time in two distinct ways:

- (i) Due to movement of the articulators vocal tract shape changes and hence the frequencies at which formant occurs also change. This variation of formants are usually slow and takes several pitch periods during speech production except during transitions in consonant vowel units.
- (ii) The variation of formant parameters within one pitch period when the articulators themselves do not move. This is because the vocal folds oscillates between open and closed phase. During closed phase the vocal tract at one end is closed and speech signal is mainly due to free resonances. But during open phase the trachea, the vocal folds and the vocal tract are acoustically coupled and this changes the free resonances. Situation is more complicated because unlike closed phase beginning of open phase is not abrupt, rather the flow of

air through the vocal folds increases initially and decreases subsequently as a function of time. So the characteristics of the vocal tract system in the open phase are not constant but signal dependent. Because of the changes within a pitch period it is necessary to determine the formant parameters separately for each of the open and closed phase regions.

1.2 Brief Description of the Problem

The speech synthesised using a model which assumes constant formant frequency in a frame, is generally very intelligible but often sounds unnatural. The formant frequencies can be taken to be constant if the assumption that the glottal source and the vocal tract are linearly separable and do not interact with each other is true. But it is found that they actually interact and this leads to variation of formant frequency even in a single pitch period. Here we are interested to track formant parameter variations with time within a single pitch period and from period to period. For changes within a pitch period small data window size are to be used to determine formant parameters separately for each of the open and closed phase regions. But to extract these informations one should have a proper knowledge about where the instants of significant excitation and secondary excitation occur which mark the beginning of closed and open phase respectively. From there one can find the durations of close and open phases which in turn give an idea of placement and size of the data window to be used to get reasonably good estimate of formant parameters. So here we make an attempt to understand the above issues by using covariance LPC analysis of speech signal.

1.3 Organisation of Chapters

Chapter 2 describes the mathematical formulations and understanding required in the background for Covariance method of linear prediction as a least square approach and glottal inverse filtering. Chapter 3 deals with a measure for locating instants of glottal closure and opening and extraction of formant from speech wave by peak picking of linear prediction spectra. Noise sensitivity in extracting formants at different S/N power ratio are also investigated. Chapter 4 describes the simulation of synthetic vowels using formant synthesiser with a given voiced source excitation model and the results of the single and multi-formant simulation. In Chapter 5 experimental results on natural speech signal collected from Timit data base for same speaker are tabulated. Scope for further work is also discussed here.

Chapter 2

LP Model Formulations

2.1 All Pole Model of Voiced Speech Production

Block diagram of Fig. 2.1 shows the speech production model for voiced speech where G(z) is glottal model filter, V(z) is vocal tract model and R(z) is radiation impedance. The pertinent signals and their z-transforms are defined by

$E(z) \leftrightarrow e(n)$ glottal excitation model signal,

$U_G(z) \leftrightarrow u_G(n)$ glottal volume velocity signal,

$U_L(z) \leftrightarrow u_L(n)$ lip volume velocity signal,

$S(z) \leftrightarrow s(n)$ speech pressure wave signal

For voiced sound e(n) is taken to be a periodic train of impulses. Vocal tract model V(z) is assumed to be an all pole model chosen as $V(z) = \frac{1}{1-\sum_{k=1}^{p} a_k z^{-k}}$ so there can be at most p/2 formants and a_k's are predictor coefficients, R(z) is the radiation impedance of lips which is a differentiating filter chosen as $R(z) = (1 - z^{-1})$. Note that R(z) is zero at zero frequency.

From Fig. 2.1

$$U_G(z)V(z)R(z) = S(z)$$

$$or, \quad U_G(z) = S(z) \cdot \frac{1}{V(z)} \cdot \frac{1}{R(z)}$$

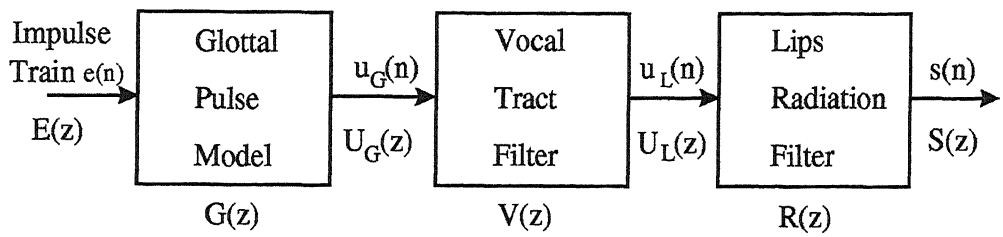


Figure 2.1: Block diagram of voiced speech production model

Our aim is to find out $U_G(z)$ from S(z). Now interchanging V(z) and R(z) in speech production model shown in Fig. 2.1 we define a driving function as

$$Q(z) = U_G(z)R(z)$$

Q(z) is applied to V(z) to form S(z). Hence speech production model assumes the form shown in Fig. 2.2. From the above equation

$$q(n) = u_G(n) \star r(n)$$

where $\star$ denotes convolution and q(n) is termed as driving function also known as voice source excitation. Hence linear model of Fig. 2.1 is equivalently described by the model in Fig. 2.2. Since the radiation term is zero at zero frequency q(n) must be a zero mean signal. From Fig. 2.1, we can write

$$\begin{aligned} S(z) &= Q(z)V(z) \\ &= \frac{Q(z)}{1 - \sum_{k=1}^{p} a_k z^{-k}} \end{aligned}$$

$$or, \qquad S(z) - (\sum_{k=1}^{p} a_k z^{-k})S(z) = Q(z)$$

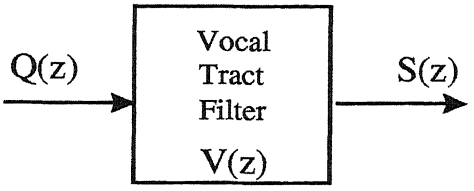


Figure 2.2: Simplified model of Fig 2.1

Taking inverse Z-transform, we get

$$s_n - \sum_{k=1}^{p} a_k s_{n-k} = q_n$$

$$s_n = \sum_{k=1}^{p} a_k s_{n-k} + q_n \tag{2.1}$$

Hence Eq. (2.1) can be interpreted as an all pole model where s_n is given as a linear combination of past p values and some input q_n which is a driving function. In the glottal close phase $q_n = 0$; hence

$$s_n = \sum_{k=1}^{p} a_k s_{n-k}$$

So for a particular signal s_n, the problem is to determine the prediction coefficients a_k's in least square sense.

2.2 Method of Least Squares

Here we assume that the input $q_n = 0$. Therefore the signal s_n can be predicted only approximately from a linearly weighted summation of past samples. Let this

approximation of s_n be $\tilde{s}_n$, where

$$\tilde{s}_n = +\sum_{k=1}^{p} a_k s_{n-k} \tag{2.2}$$

The error between the actual value s_n and the predicted value $\tilde{s}_n$ is given by

$$e_n = s_n - \tilde{s}_n \ = s_n - \sum_{k=1}^{p} a_k s_{n-k}$$

where e_n is known as residual. In the method of least squares the parameters a_k are obtained as a result of minimisation of the mean of total squared error with respect to each of the parameters. Average prediction error from short time segment [p+1,N] is given by

$$\begin{aligned} E_n &= \sum_{n=p+1}^{N} e_n^2 \\ &= \sum_{n=p+1}^{N} (s_n - \tilde{s_n})^2 \end{aligned}$$

$$or, \quad E_n = \sum_{n=p+1}^{N} (s_n - \sum_{k=1}^{p} a_k s_{n-k})^2 \tag{2.3}$$

Differentiating E w.r.t a_i and putting $\frac{dE_n}{da_i} = 0$, $i = 1, 2, \ldots, p$, we get

$$2 \sum_{n=p+1}^{N} (s_n - \sum_{k=1}^{p} a_k s_{n-k}) s_{n-i} = 0$$

$$or, \quad \sum_{n=p+1}^{N} s_{n-i} s_n - \sum_{k=1}^{p} a_k \sum_{n=p+1}^{N} s_{n-i} s_{n-k} = 0$$

$$or, \quad \phi_{i,0} - \sum_{k=1}^{p} a_k \phi_{i,k} = 0$$

$$or, \quad \sum_{k=1}^{p} a_k \phi_{i,k} = \phi_{i,0} \tag{2.4}$$

$$where, \quad \phi_{i,k} = \sum_{n=p+1}^{N} s_{n-i} s_{n-k} \tag{2.5}$$

$$i = 1, 2, \ldots, p; \qquad k = 0, 1, 2, \ldots, p$$

From Eq. (2.5),

$$\phi_{i,0} = \sum_{n=p+1}^{N} s_{n-i}s_n \quad and \quad \phi_{0,0} = \sum_{n=p+1}^{N} s_n^2$$

From Eq. (2.3),

$$E_n = \sum_{n=p+1}^{N} s_n^2 - \sum_{k=1}^{p} a_k s_n s_{n-k}$$

$$or, \quad E_n = \phi_{0,0} - \sum_{k=1}^{p} a_k \phi_{0,k} \tag{2.6}$$

Eq. (2.4) can be written in matrix form

$$\Phi\, a \;=\; \Psi \tag{2.7}$$

This is called normal equation.

2.3 Cholesky Decomposition

The predictor coefficients a_k, $1 \leq$ k $\leq$ p can be computed by solving a set of p equations with p unknowns from normal Eq. (2.7). We note from the Eq. (2.7) that matrix of coefficients is a covariance matrix Φ is symmetric and in general positive semidefinite. Hence Eq. (2.7) can be solved more efficiently by squareroot or Cholesky decomposition method. Numerical stability properties of Cholesky method is considered to be quite stable [10]. Matrix Φ in Eq. (2.7) can be expressed as

$$\Phi = VV^T \tag{2.8}$$

$$\phi_{ij} = \sum_{k=1}^{i} v_{ik}v_{jk}, \qquad i \leq j$$

where $V = (v_{ij})$ is a lower triangular matrix and $v_{ji} = 0$ for $i > j$.

$$\phi_{ij} = \sum_{k=1}^{i-1} v_{ik}v_{jk} + v_{ii}v_{ji}$$

$$or, \quad v_{ii}v_{ji} = \phi_{ij} - \sum_{k=1}^{i-1} v_{ik}v_{jk}, \qquad j \geq i \geq 1$$

$$or, \qquad v_{ji} = \frac{\phi_{ij} - \sum_{k=1}^{i-1} v_{ik} v_{jk}}{v_{ii}} \tag{2.9}$$

where, $\quad i = 1, 2, \ldots, p \quad$ and $\quad j = i+1, \ldots, p$

v_{ii} can be obtained by setting diagonal elements $i = j$.

$$\begin{aligned} \phi_{ii} &= \sum_{k=1}^{i} v_{ik}^2 \\ &= \sum_{k=1}^{i-1} v_{ik}^2 + v_{ii}^2 \end{aligned}$$

$$or, \qquad v_{ii}^2 = \phi_{ii} - \sum_{k=1}^{i-1} v_{ik}^2 \tag{2.10}$$

Therefore, $\quad v_{11}^2 = \phi_{11}$

From Eqs. (2.7) and (2.8),

$$VV^T a = \Phi\, a = \Psi$$

$$or, \qquad VY = \Psi \tag{2.11}$$

$$where \quad V^T\, a = Y \tag{2.12}$$

From Eq. (2.11),

$$\sum_{j=1}^{i} v_{ij} y_j = \psi_i$$

Solving for j = i for y_i, we get

$$\sum_{j=1}^{i-1} v_{ij} y_j + v_{ii} y_i = \psi_i$$

$$or, \qquad y_i = \frac{(\psi_i - \sum_{j=1}^{i-1} v_{ij} y_j)}{v_{ii}} \tag{2.13}$$

where p $\geq$ i $\geq$ 2 with initial condition $y_1 = \frac{\psi_1}{v_{11}}$ Having solved for y equation 2.12 can be solved for finding a. From Eq. (2.12)

$$\sum_{j=i}^{p} v_{ji} a_j = y_i$$

$$or, \qquad \sum_{j=i+1}^{p} v_{ji} a_j + v_{ii} a_i = y_i$$

$$or, \quad a_i = \frac{y_i - \sum_{j=i+1}^{p} v_{ji} a_j}{v_{ii}}$$

where 1 ≤ i ≤ p-1 with initial condition $a_p = \frac{y_p}{v_{pp}}$

From Eq. (2.7)

$$\begin{aligned} E_n &= \phi_{0,0} - \sum_{k=1}^{p} a_k \phi_{0,k} \\ &= \phi_{0,0} - a^T \psi \\ &= \phi_{0,0} - Y^T V^{-1} \psi \\ &= \phi_{0,0} - Y^T Y \\ &= \phi_{0,0} - \sum_{j=1}^{p} y_j^2 \end{aligned} \tag{2.14}$$

From Eqs. (2.5) and (2.14) normalized prediction error

$$\eta = \frac{E_n}{\phi_{0,0}} = 1 - \frac{\sum_{j=1}^{p} y_j^2}{\phi_{0,0}} \tag{2.15}$$

2.4 Log Determinant Measure

The determinant of the $(p+1) \times (p+1)$ autocovariance matrix Φ is calculated over the analysis window interval [p+1, N]. Hence for the determinant

$$\begin{aligned} det\Phi &= (detV)^2 \\ &= \Pi_{i=1}^{p} v_{ii}^2 \\ &= v_{11}^2 v_{22}^2 \ldots v_{pp}^2 \end{aligned}$$

$$or, \quad logdet\Phi = \sum_{i=1}^{p} logv_{ii}^2 \tag{2.16}$$

where v_{ii} can be obtained from Eq. (2.10)

2.5 Glottal Inverse Filter

Referring to Fig. 2.3, we have

$V(z) = \frac{1}{1+\sum_{k=1}^{p} a_k z^{-k}}$ and $R(z) = 1 - z^{-1}$

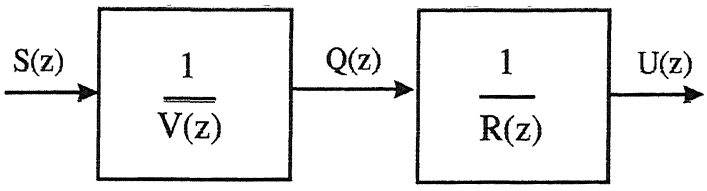


Figure 2.3: Glottal inverse filtering model

Having found out a_k's by covariance analysis speech signal is passed through inverse filter V(z) to obtain driving function or voice source excitation which is then integrated 1/R(z) to obtain glottal volume velocity (g-v-v) waveform. Note an absolute dc value cannot be obtained for U(z) since R(z) is zero at zero frequency.

2.6 Linear Predictive Spectral Matching

Insights gained from the frequency domain analysis will lead to new applications for linear predictive analysis. The error e_n between the actual signal and the predictive signal is

$$e_n = s_n - \sum_{k=1}^{p} a_k s_{n-k}$$

Applying z-transform,we get

$$E(z) = (1 - \sum_{k=1}^{p} a_k z^{-k}) S(z)$$

$$E(z) = A(z)S(z) \tag{2.17}$$

where A(z) is the inverse filter. Assume a deterministic signal s_n, and applying Perseval's theorem, the total error to be minimised is given by

$$E = \sum_{n=-\infty}^{+\infty} e_n^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |E(e^{jw})|^2 dw \tag{2.18}$$

where $E(e^{jw})$ is obtained by evaluating E(z) on unit circle $z = e^{jw}$. Denoting the power spectrum of the signal s_n by $P(w)$, where

$$P(w) = |S(e^{jw})|^2 \tag{2.19}$$

we have from above equations

$$E = \frac{1}{2\pi}\int_{-\pi}^{\pi} P(w)A(e^{jw})A(e^{-jw})dw$$

Now we will investigate in what way signal spectrum $P(w)$ is approximated by the all pole model spectrum $\hat{P}(w)$ in autocovariance LP analysis given by

$$\begin{aligned}
\hat{P}(w) &= |H(e^{jw})|^2 \\
&= \frac{G}{|A(e^{jw})|^2} \\
&= \frac{1}{|A(e^{jw})|^2}, \quad \textit{assuming gain } G = 1 \\
&= \frac{1}{|1 + \sum_{k=1}^{p} a_k e^{-jkw}|^2}
\end{aligned}$$

From Eq. (2.17),

$$|E(e^{jw})|^2 = |A(e^{jw})|^2 |S(e^{jw})|^2$$
$$\textit{or,} \quad P(w) = \frac{|E(e^{jw})|^2}{|A(e^{jw})|^2}$$

But $\hat{P}(w) = \frac{1}{|A(e^{jw})|^2}$, hence we see that if $P(w)$ is being modelled by $\hat{P}(w)$, then error power spectrum $|E(e^{jw})|^2$ is being modelled by a flat spectrum equal to 1. From Eq. (2.18), total error

$$E = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{P(w)dw}{\hat{P}(w)} \tag{2.20}$$

Minimising total error E is equivalent to the minimisation of the integrated ratio of the signal spectrum $P(w)$ to its approximation $\hat{P}(w)$. Hence we restate the problem of linear prediction as follows: Given some spectrum $P(w)$ we wish to model it by $\hat{P}(w)$ such that the integrated ratio between the two spectra as in Eq. (2.20) is minimised. Increasing the value of the order of the model 'p' resulting in a better fit of $\hat{P}(w)$ to $P(w)$. In the limit p $\rightarrow \infty$ two spectra becomes identical i.e.

$$\hat{P}(w) = P(w) \qquad as \;\; p \rightarrow \infty \tag{2.21}$$

This statement says that we can approximate any spectrum arbitrarily closely by an all pole model. Eq. (2.21) can be written alternatively $|H(e^{jw})|^2 = |S(e^{jw})|^2$ as p$\rightarrow \infty$. But it is not necessarily true that $H(e^{jw}) = S(e^{jw})$ i.e. the frequency response of the model need not equal to the Fourier Transform of the signal. This is so because $S(e^{jw})$ need not be minimum phase, where as $H(e^{jw})$ is required to be minimum phase since it is the transfer function of an all pole filter with poles inside the unit circle. An important point to note that since we assume the availability of the signal spectrum $P(w)$, any desired frequency shaping or scaling can be performed directly on the signal spectrum before linear predictive modelling is applied. The LP error measure E in Eq. (2.20) has two major properties, (i) global and (ii) local

Global Property

Because the contributions to the total error are determined by the ratio of the two spectra, the matching process should perform uniformly over the whole frequency range. It makes sure that spectral match at frequencies with little energy is just as good as the match at frequencies with high energy.

Local Property

This property deals with how the match is done in each small region of the spectrum. If $E(w) = P(w)/\hat{P}(w)$ then from Eq. (2.20)

$$E = \frac{1}{2\pi}\int_{-\pi}^{\pi} E(w)dw = 1 \qquad \text{(when match occurs)}$$

$E(w)$ can be interpreted as the 'instantaneous error' between $P(w)$ and $\hat{P}(w)$ at frequency w. Above equation can be interpreted as that there are values of $E(w)$ greater and less than 1 such that average value is equal to 1. In terms of two spectra this means that $P(w)$ will be greater than $\hat{P}(w)$ in some regions and less in others such that above equations applies.However the contribution to the total error is more significant when $P(w)$ is greater than $\hat{P}(w)$ than when $P(w)$ is smaller (e.g. $E(w) = 2$ contributes more to total error than a ratio of 1/2).

If we see LP spectrum we observed that a significant feature of the LP spectrum matches the signal spectrum much more closely in the regions of large energy (i.e. near spectrum peaks) than near the regions of low signal energy (i.e. near spectrum valleys). This is expected in view of Eq. (2.20) since regions where $|S_n(e^{jw})| > |H(e^{jw})|$ contribute more to the total error than regions where $|S_n(e^{jw})| < |H(e^{jw})|$. Thus the LP spectral error criterion favours good fit the spectral peaks, whereas fit near the spectral valleys is nowhere near as good.

2.7 Considerations in the Choice of Analysis Parameters

There is a relationship between choice of filter length and the accuracy of the resonance to be estimated. The order 'p' of the linear predictive analysis can effectively control the degree of smoothness of the resulting spectrum. Since our objective is to obtain a representation of only the spectral effects of the glottal pulse, vocal tract and radiation, it is clear that we should choose 'p' so that the formant resonances and the general spectrum shape are preserved. Fortunately it has been discovered that predictor order 'p' is not a strong function of the particular speech sound. However, it is a strong function of system sampling rate f_s. A rule of thumb [6] is that for $6 \leq f_s \leq 18$ Hz the equation $p = f_s + \gamma$, where γ =2 to 5 has been

found generally sufficient for the analysis. For example f_s=16 kHz p=18 for $\gamma = 2$. Physical interpretation of this result is simply that independent of the sampling rate, roughly one complex pole pair is required to span approximately every 900 Hz. With higher value of 'p' resolution of peak will increase.But as 'p' is increased more and more, better and better approximation to input spectrum as opposed to the resonance behaviour is obtained. It has been observed that $p = 18$ for f_s=16 kHz quite reasonable estimates of the formant frequencies are possible by simple peak picking.

2.8 Choice of Autocovariance over Autocorrelation Method

A question always arises as to whether to use the autocorrelation method or covariance method in estimating the prediction parameters. The covariance method is quite general and can be used with no restrictions. The only problem is that of the stability of the resulting inverse filter which is not a severe problem generally. In autocorrelation method, on the other hand the filter is guaranteed to be stable, but problems of parameter accuracy can arise because of the necessity of windowing (truncating) the time signal. This is usually a problem if the signal is a portion of an impulse response (note in speech glottal closing and opening is excited by some kind of impulse). For example if the impulse response of an all pole filter is analysed by covariance method, the filter parameters can be computed accurately from only a finite number of samples of the signal (i.e. it is helpful to analyse speech parameters within small finite glottal closing and opening). Using autocorrelation method one cannot obtain the exact parameter values unless the whole infinite response is used in the analysis. However in practice, very good approximations can be obtained by truncating the impulse response at a point where most of the decay of the response has already occurred.

Chapter 3

Extraction of the Instants of Glottal Closure/Opening and Formants from Speech Signal

3.1 Using the Normalized Prediction Error

Variation of normalized prediction error as a function of position of the analysis frame within a single stationary speech segment is analysed. The magnitude of this prediction error variation depends on the analysis frame size (i.e. number of samples contained within a frame). However maxima of the prediction error do not coincide with the instants of closure. A more specific investigation of the error as a function of window position shows p-error jumps up when the analysis window from complete closed phase entered into open phase or the reverse case i.e. analysis window from completely within open phase encounters instants of significant excitation, the mark of the beginning of closed phase. Prediction error is low when time derivative of the glottal pulse is zero. When window totally enters inside the closed phase there is a sudden drop of prediction error. This is the point of glottal closure. Since the transition to closed phase is much more abrupt than

transition to open phase normalized prediction error η changes much more rapidly at the point of closure. However by preemphasising the speech data before computing the covariance analysis, the shape of the opening phase can becomes steeper and is more accurately determined from η. The point of glottal closure is given by the sample location when $\eta < \eta_{th}$. There is no theoretical criterion for defining threshold normalized error η_{th}. In practice η_{th} is obtained as a value just exceeding sample $\eta(n)$ that is abruptly lower than the preceding sample. Normalized p-error signal in Fig. 3.1 shows its behaviour near glottal closure and open for any arbitrary vowel waveform. t1(c) and t5(c) indicates the points of glottal closures where the fall of prediction error is sharp. Note the exact location of glottal open is difficult to obtain from this p-error waveform as change in error is gradual there.

This method of extracting glottal closure as suggested in paper [4] seems to have been in problem for smaller analysis window length when prediction error becomes uncorrelated and abrupt changes of normalized p-error occurs at several places which sometimes wrongly interpreted as glottal closure.

3.2 Using Autocovariance Determinant

Strube [7] made a hypothesis that the determinant of the autocovariance matrix is maximum if the beginning of the window interval on which autocovariance matrix is calculated coincides with the glottal closure. This is verified by comparison with the prediction error method. The value of the determinant is not only a measure for the linear dependence, but is strongly influenced by the signal amplitude. This measure had been suggested by Strube for locating the instants of glottal closure and is extended to find glottal open by preemphasizing the speech signal before covariance analysis. It is seen that that its peaks are within a few points from the point of closure predicted by η measure. However theoretically there is no precise mathematical relationship between the point of closure and Log determinant peak.

From experiment it has been found out that shorter analysis window less than closed phase and open phase duration must be used. Longer analysis window produce Log determinant values with flat tops.

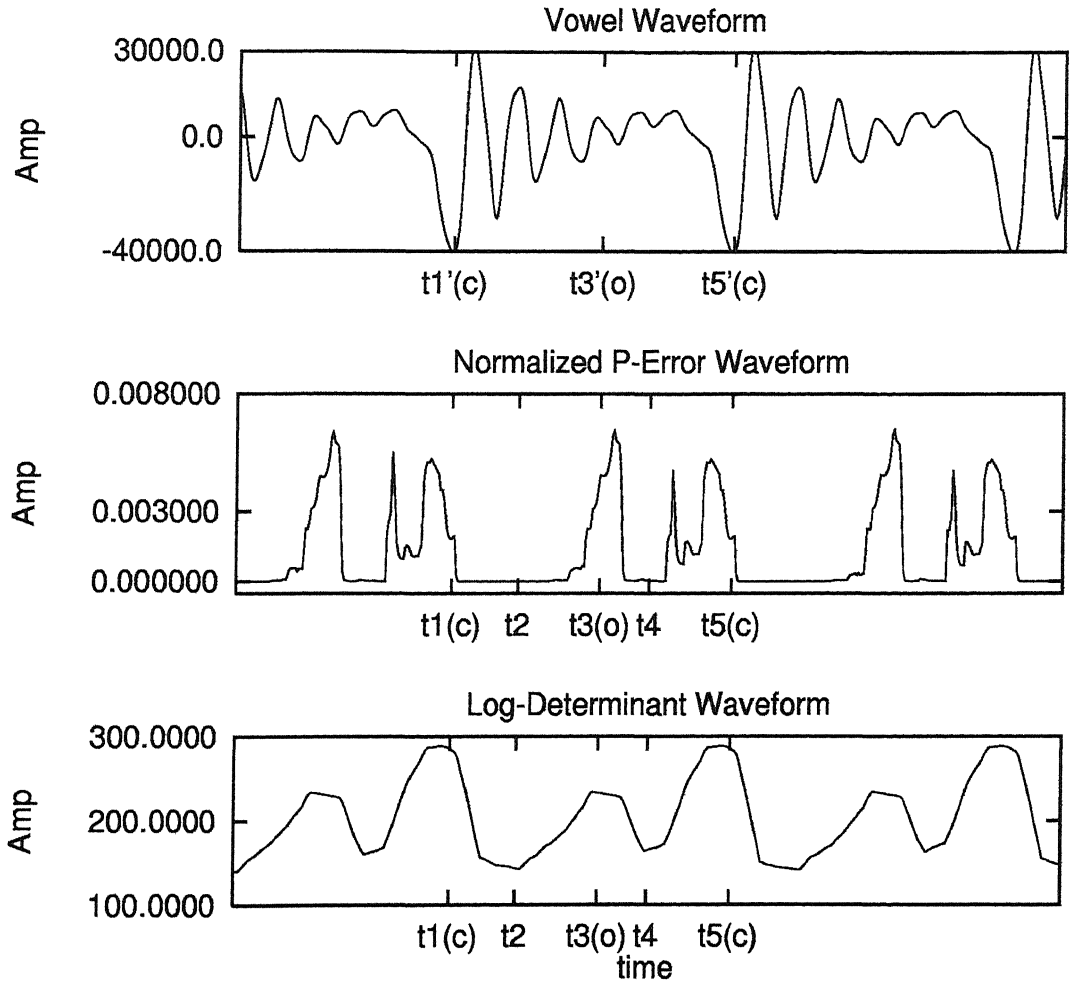


Figure 3.1: Example of Normalized p-error and Log-det waveform

Accuracy of the measure depend on strong peaks in speech wave generated by sharp glottal closures and strong secondary excitation at glottal open and not so much on how well the speech fits the linear all pole model. Log-determinant waveform in Fig. 3.1 shows its behaviour.

Interpretation of Log determinant waveform:

Referring to Fig. 3.1 it is noticed from stationary vowel waveform that in closed phase output signal decays approximately exponentially, therefore the logarithm of the determinant is linearly decreasing with time in the interval t1 to t2. As the window enters in the open phase a nonzero excitation occurs and the determinant will go up strongly at t3. Again the impulse response in the open phase decreases with time until another nonzero excitation at the beginning of the closed phase (called the instant of significant excitation) occurs when determinant will go up strongly at t5. After that determinant again decreases as prediction improves. Thus overall curve exhibit sawtooth like shape. Interval t5 to t3 gives approximately open phase duration including the interval in which significant excitation occurs and t3 to t1 gives approximately the close phase duration including the secondary excitation at the beginning of open phase. We also note that secondary excitation is weaker than significant or main excitation. Another thing to be noted that maximum of the curve indicates the end of the excitation. It may happen that due to weak secondary excitation peak at 'o' may not be properly visible. It is seen that t2 is much lower than t4 indicates that at t2 prediction error is small or prediction improves there.

3.3 Extraction of Formant Trajectory

(i) Formant Frequency and Amplitude

This is based upon digital inverse filter formulation and is quite accurate for estimating resonances or formant structure of voiced speech. Predictor coefficients a_k's are obtained from covariance analysis of speech waveform. Then LP spectrum is obtained by evaluating the magnitude of the transfer function $H(z)$ of the filter represented by the coefficients a_k's at N_DFT equally spaced samples along the unit circle in z- plane. N_DFT is equal to the number of DFT points to be taken.

$$H(z) = \frac{a_o}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{3.1}$$

where $H(z)$ is evaluated at $z = exp(j2\pi n/N\_DFT)$ for n = 0,1,2,3,...,$(N\_DFT - 1)$ and gain a_o is assumed to be unity for simplification.Therefore,

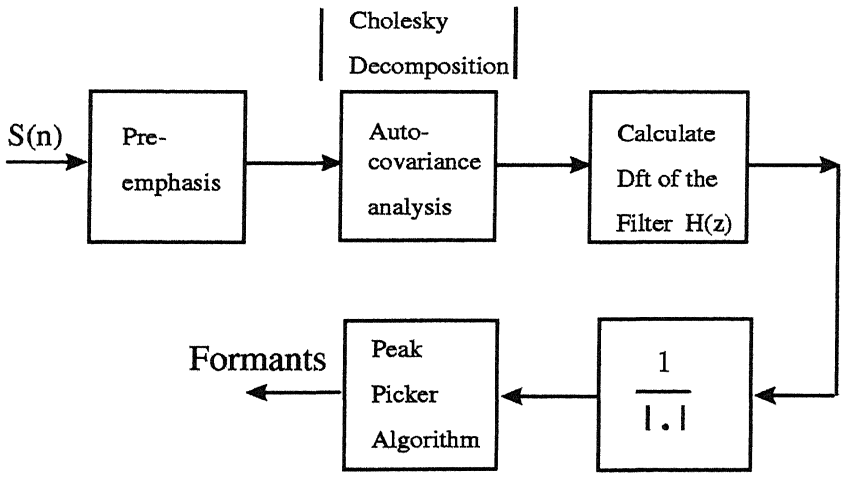


Figure 3.2: Block diagram to find formant tracks from LP magnitude spectra

$$H(e^{jw}) = \frac{1}{1 - \sum_{k=1}^{p} a_k e^{-\frac{j2\pi nk}{N\_DFT}}}$$
$$= \frac{1}{1 - \sum_{k=1}^{p} a_k cos(\frac{2\pi nk}{N\_DFT}) + j \sum_{k=1}^{p} a_k sin(\frac{2\pi nk}{N\_DFT})}$$

From the above equation LP amplitude and phase spectra can be obtained. N_DFT value can be chosen arbitrarily large to increase frequency resolution, at the expense of computing time. In our experiment we choose $N\_DFT = 8192$, resulting an approximately 2 Hz spectral resolution. Now formants are obtained by simple peak picking of LP amplitude spectra. In peak picking peaks are selected from a smoothed linear prediction spectrum at each frame, sample location of peaks multiplied by the scale factor gives the first three formants frequencies. The first three formants can be uniquely defined as the first three peaks in the reciprocal of the inverse filter spectrum. Details are described in block diagram of Fig. 3.2. Preemphasis is done before autocovariance to emphasise the higher formants due to windowing and also to de- emphasise the low frequency non formant peaks occasionally caused by strong fundamental frequency component.

(ii) Formant Bandwidth

To track the bandwidth roots $\rho_k e^{j\theta_k}$ of the predictor polynomial $A(z) = a_o + a_1 z^{-1} + \ldots + a_p z^{-p}$ are found out in each analysis frame which is shifted as a function of position of analysis frame. Each formant is a free resonance of the vocal-tract system and thus corresponding time signal can be written as a sum of complex resonances.

$$r(n) = \sum_{k=1}^{p} A_k \rho_k^n e^{j\theta_k n}$$

$$or, \quad r(n) = \sum_{l=1}^{p/2} \rho_l^n (A_l e^{j\theta_l n} + \tilde{A}_l e^{-j\theta_l n}) \tag{3.2}$$

Where n is the time index, and number of formants below $f_s/2$ are $p/2$ and k is the index of particular formant,θ_k is the normalised formant frequency, $\pi < \theta < \pi$, and ρ_k determines formant bandwidth or damping $0 < \rho < 1$ and A_k is the complex

formant amplitude. Since r(n) is real formant, from Eq. (3.2) resonances occur in complex conjugate pairs. The formant frequency F_k and bandwidth B_k in hertz are given by

$$\begin{aligned} F_k &= \frac{fs}{2\pi}\theta_k \\ and \quad B_k &= \frac{-fs}{\pi} ln(\rho_k) \end{aligned}$$

It is found that some of the roots would be real and rest would be complex pole pairs which might or might not be formants. The extra roots generally (not always) have large bandwidths and thus not detected by moving an analysis contour only along the unit circle in the z-plane. Now out of those pole pairs, one would have to select three on the basis of frequency location sufficiently narrow bandwidth to be the first three formants. It is often found that corresponding to some of the roots spurious formants come out whose bandwidths differ considerably. Hence some kind of formant continuity criterion is used to pick up corresponding bandwidth. This is one of the reason why polynomial root solving method is avoided to find out formants, in addition it takes more computation time.

3.4 Noise Sensitivity

We study the effect of white Gaussian noise on the estimate of formant frequencies. Since placing of window also plays an important role in finding out correct formant frequencies, experiment boils down to extraction of instants of significant excitation under the influence of noise. Since higher formants contain less energy, these formants are affected more by the noise and our study is to see at what signal to noise ratio LPC breaks down. It is mentioned in paper [2] that $S/N = 40dB$ is the limit for LPC to breakdown. All these issues have been discussed in chapter 4.

Chapter 4

Simulation Results

4.1 Formant Synthesis

Vowels are synthesised using digital resonators acting as vocal tract filter. It gives an approximation to a speech waveform by a simple set of rules formulated in acoustic domain. This does not take into account articulatory motion of the vocal tract. Two parameters are used to specify the input-output characteristics of a resonator, the resonant (formant) frequency F_k and the resonance bandwidth B_k. Output samples of the resonator $y(nT)$, are computed from the input sequence $x(nT)$ by the equation

$$y(nT) = Ax(nT) + By(nT - T) + Cy(nT - 2T) \tag{4.1}$$

where $y(nT - T)$ and $y(nT - 2T)$ are previous two sample values of the output sequence $y(nT)$.Constants A,B and C related to resonant frequency F_k and bandwidth B_k of the resonator are given by [3]

$$C = -exp\left(\frac{-2\pi B_k}{f_s}\right) = -r_k \star r_k \qquad \text{where } r_k = exp(\frac{-\pi B_k}{f_s})$$
$$B = 2r_k cos\left(\frac{-2\pi F_k}{f_s}\right) = 2r_k cos\theta_k \qquad \text{where } \theta_k = \frac{2\pi F_k}{f_s}$$

$$and \quad A = (1 - B - C) = 1 - 2r_k cos\theta_k - r_k^2$$

We have chosen 16 kHz as sampling frequency as most of the sound energy of speech is contained in frequency between 80 Hz to 8000 Hz. Hence 16 samples corresponding to 1 ms duration. Formants and bandwidth can be estimated from the equation

$$F_k = \frac{fs}{2\pi}\theta_k \tag{4.2}$$

$$and \quad B_k = -\frac{fs}{\pi}ln(r_k) \tag{4.3}$$

4.2 Voiced Source Excitation Model

For simulation purpose we assume a model which gives some kind of dynamical simulation of vocal fold oscillations of a two mass vocal fold model. Voice source excitation model can be thought of as equivalent to differentiation of glottal vol. velocity at lips. Childers et al. [1] demonstrated that simulating source-tract interaction can improve the quality of synthetic speech. Our model of speech production shown in Fig. 4.1, models the skewness of the glottal volume velocity waveform using Liljencrants-Fant (LF) model [1]. The LF model uses some parameters to model the differentiated glottal volume velocity. The model consists of two parts. The first part t_o to t_e is an exponentially growing sinusoid represented by

$$\frac{dU_g(t)}{dt} = E(t) = E_o\ \dot{e}^{\alpha t} sin(w_g t)$$

This portion of the model represents the volume velocity from the opening of the glottis to the instant at which the main excitation occurs which is also the instant at which the maximum discontinuity in the glottal volume -velocity occurs. This discontinuity normally coincides with the instant of the maximum negative derivative.

The three parameters pertaining to the first segment of the LF model have the following interpretation:

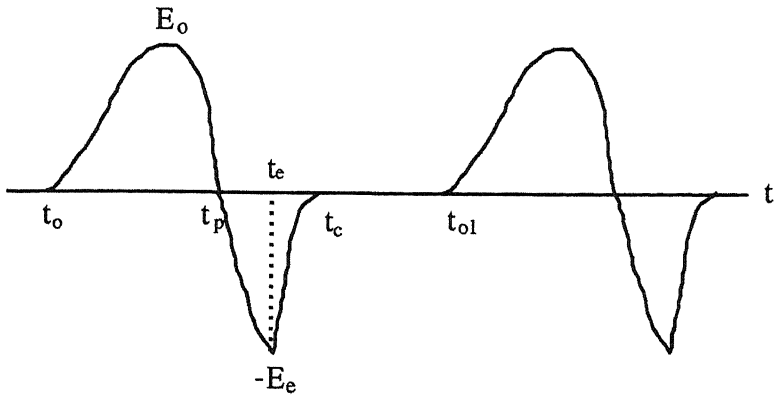


Figure 4.1: LF model of the differentiated glottal volume velocity waveform

(1) E_o is the scale factor.

(2) α governs the exponentially growing amplitude

(3) $w_g = \pi/t_p$where t_p is time from glottal opening to maximum glottal volume velocity which is the integral of the above differentiated g-v-v.

The second part of the model t_e to t_c called return phase is an exponential segment that allows residual volume velocity to come to zero following the main discontinuity when the vocal folds close. Model equations have been simulated in Difference equation of the Hypersignal DSP software. Measurments are also taken in Hypersignal DSP package. We have chosen $E_o = 500, \alpha = 0.03, w_g = 0.0145$ to give a practical shape of differentiated g-v-v waveform. It can be seen that the effect of glottal pulse in the frequency domain is to introduce low-pass filtering effect. Longer the return phase the lower the cutoff frequency. We kept this glottal excitation model fixed throughout the experiment and designated it as LF (Liljencrants-Fant) excitation. There is an additional requirment that the area of the differentiated LF model must be zero. This condition maintains a constant baseline for the volume velocity.

4.3 Construction of Synthetic Vowels

Simulated digital resonator having parameters set for the first formant is excited by periodic voice source pulse shown in Fig. 4.1. Then the resonator parameters are changed corresponding to second formant frequency and bandwidth. Output of the digital resonator already obtained is again applied to its input. The same is repeated for the third formant also. Thus the final output is the required synthetic vowel. We simulated vowels with the first three formants because the frequencies of the lowest three formants vary substantially with changes to articulation. The frequencies and bandwidth of the fourth and fifth formants do not vary as much and can be held constant. These higher frequency resonators help to shape the overall spectrum but otherwise contribute little to intelligibility of vowels. Also bandwidth vary very little so that all formant bandwidths might be held constant in some applications in which case only F_1, F_2 and F_3 would be varied to simulate the vocal tract transfer functions for non-nasalised vowels. Therefore all the parameters namely, formant frequency, bandwidth and amplitude of the synthetic vowel formed by the digital resonator are known. We simulate a set of signals to study the performance of our formant frequency estimation methods.

4.4 Single-formant Simulation

(i)Estimated frequency under time variation of formant frequency:

We simulated synthetic vowel whose formant varies linearly with time with the help of formant synthesizer excited by the given voice source LF model shown in Fig. 4.1. True formants set in the formant synthesizer for each transition cases are $(1000 + 0.5 * n), (1000 + 1.25 * n), (1000 + 2.0 * n)$ Hz respectively where n is the sample number. Initial value of the formant chosen is 1000 because we are interested

to see the formant transitional behaviour in the frequency region between second and first formants. Predictor order p is 10 and data window length N is 42 hence analysis window length is 32. Since sampling frequency is 16 kHz, analysis window length is equivalent to 2 ms. So formant transition per sample in each case is 0.5 Hz, 1.25 Hz and 2.0 Hz respectively.

We have estimated formant frequency at three locations in closed phase and three locations in open phase of the voice source excitation. It has been observed that best location for estimating formant in open phase is corresponding to position where time derivative of the voice source excitation is zero which also is the position where prediction error is minimum. Another condition which must also be satisfied that window must be inside the open or closed phase. Referring Table 4.1, center column of each of the closed and open phase estimates namely m_{cp} and m_{op} respectively meets the above criterion and is called the "best location" to take estimate. m_{cp} is corresponding to place where window is placed in the center of closed phase and $m_{cp}-16$ and $m_{cp}+16$ are corresponding to positions 16 sample away from either side of m_{cp}. Similarly for the case of m_{op}. Values shown in Table 4.1 are taken within one pitch period of 170 samples. If we move from period to period and observe similar locations in adjacent pitch periods we see transition rate is properly maintained. It is also observed that when the formant transition rate increases estimated value deviates more from the true value. Fig. 4.2 shows formant frequency tracks at different transition rates per sample.

The effect of the predictor order: Single formant is represented by only one complex pole pair. Hence theoretically predictor order two should be sufficient to estimate the formant frequency. With this lower predictor order such as p=2 estimated values in the closed phase are very smooth near to true value but open phase estimates are far from the true value as two samples are not adequate to predict in the presence of the excitation. With the increase of predictor order it is observed that accuracy in the open phase estimates increases but at the same time fluctuation

of the estimated value in the closed phase contiguous to best window location go up slightly. This is because higher predictor order represents better approximation to input spectrum as opposed to resonance behaviour.Hence some compromise is made and p=10 is the minimum predictor order chosen to get reasonably good estimate from open phase as well as from closed phase.

(ii)Root mean square error in frequency to estimate formant from signal with additive white gaussian noise:

White Gaussian noise is added in different proportions with synthetic clean signal to simulate noisy signal. This noisy signal is given to the input of the inverse filter algorithm to find contineously varying formant frequency. This estimated value is subtracted from the true value to find estimated r.m.s frequency error at different signal to noise ratio. Measurment is done over 60 samples from three consecutive pitch periods having 20 samples from each pitch period in closed phase. Table 4.2 shows less the SNR larger is the r.m.s frequency error as expected. The glottal volume velocity is recovered by inverse filtering even from signal having SNR = 5 dB as location of the instant of significant excitation was known.

Table 4.1: Estimated frequency under time variation of formant frequency

(Predictor Order = 10; Data window size = 42)

Formant transition per sample	Formant	Estimated formant frequency in Hz					
		Samp.loc.at c.p.			Samp.loc.at o.p.		
		296	312	328	370	386	402
		$m_{cp}-16$	m_{cp}	$m_{cp}+16$	$m_{op}-16$	m_{op}	$m_{op}+16$
0.0 Hz	f_{true}	1000	1000	1000	1000	1000	1000
	f_{estlp}	998	1000	996	988	998	992
	f_{dev}	+2	0	+4	+12	+2	+8
0.5 Hz	f_{true}	1148	1156	1164	1185	1193	1201
	f_{estlp}	1141	1158	1160	1174	1190	1189
	f_{dev}	+7	-2	+4	+11	+3	+12
1.25 Hz	f_{true}	1370	1390	1410	1463	1483	1503
	f_{estlp}	1382	1392	1430	1441	1479	1480
	f_{dev}	-12	-2	-20	+22	+4	+23
2.00 Hz	f_{true}	1592	1624	1656	1740	1772	1804
	f_{estlp}	1613	1629	1676	1713	1760	1777
	f_{dev}	-21	-5	-20	+27	+12	+27

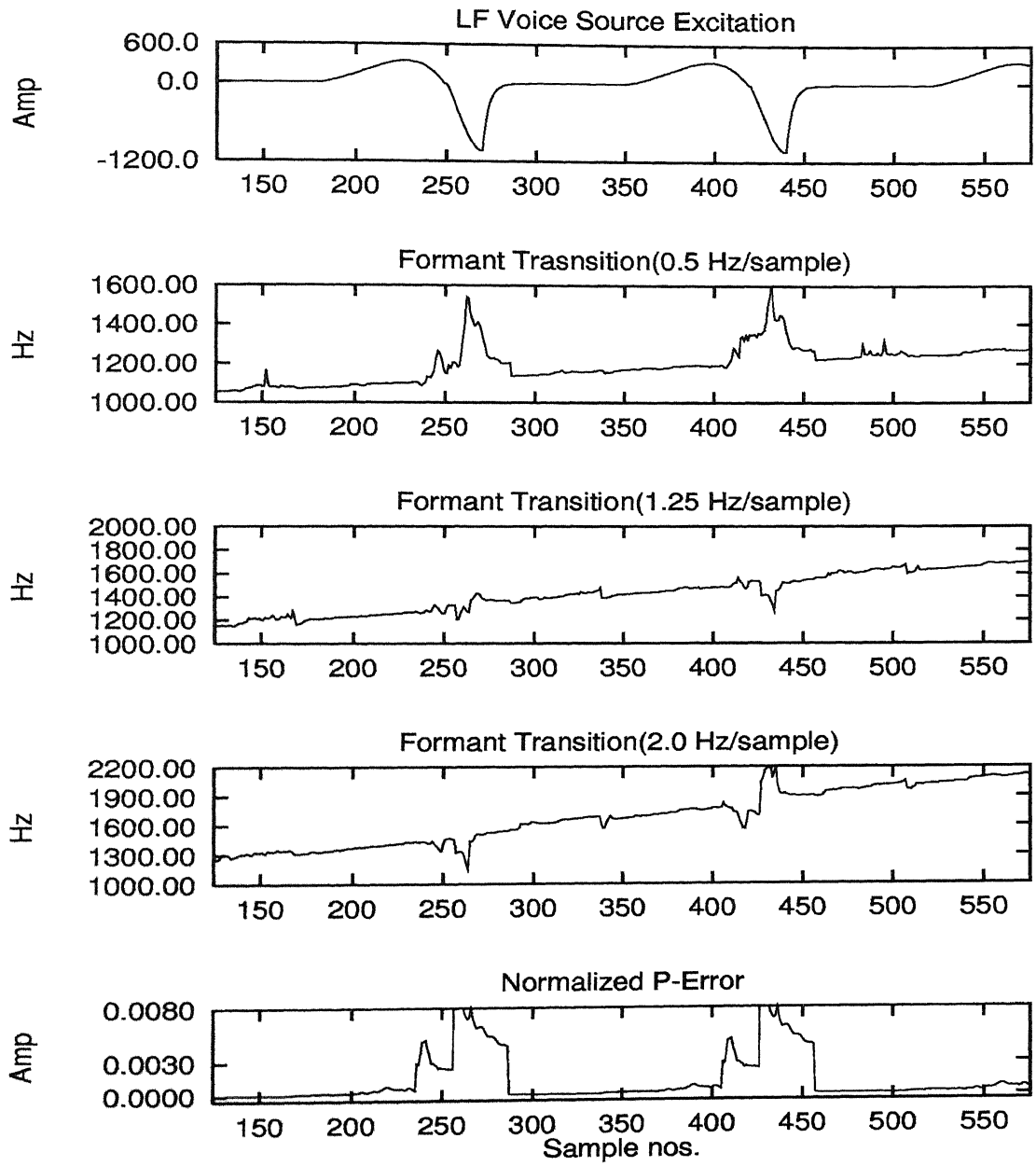


Figure 4.2: Time variation of Formant frequency

Table 4.2: Root mean square error in frequency to estimate formant from single formant signal with additive white gaussian noise

(Predictor Order = 10; Data window size = 50)

Voice Source Excitation	Single Formant	S/N Power	r.m.s. error in closed phase
LF	1500Hz	25 dB	4.09
		15 dB	13.17
		5 dB	40.31

4.5 Multi-formant Simulation

True formant frequencies 640,1500 and 2400 Hz are set during construction of synthetic vowel 'ae' in the closed phase. To distinguish open phase formants from closed phase we kept open phase formant frequency 60 Hz more than closed phase formant frequency. Bandwidth is set to 100 Hz both in closed and open phase.LF model is used as voice source excitation.

Fig. 4.3 shows formant frequency track for simulated vowel 'ae' after inverse filtering. Predictor order chosen is 18. Data window length is 50. We see from Fig. 4.3 normalized p-error is minimum when time derivative of voice source excitation is zero. Formant frequencies are estimated at those positions where p-error is minimum. Fig. 4.4 shows how Log-determinant waveform indicates start of glottal

closure and open. Glottal volume velocity is also extracted by inverse filtering from simulated vowel waveform 'ae'.

(i)Formant estimation of vowel 'ae' having different pitch period with various data window lengths:

In simulation we have chosen formant spacing uniformly to eliminate any bias due to interformant spacing. Predictor order chosen is 18. Data window lengths are 50 and 40 respectively. For pitch period of 170 samples, referring to Table 4.3, following observations are made:

- In closed phase estimation of all the three formant frequencies are very near to true value at and around the best location $m_{cp} = 470$.
- In open phase best location to estimate formant frequencies is at $m_{op} = 544$ where p-error is minimum and window is placed in the center of the open phase. Estimated frequency of 1st formant deviates more than higher formants from their true values. This may be due to the reason that in case of low frequency data window contains only 1 to 1.5 cycles which is not sufficient to estimate properly in the presence of excitation. Whereas in case of high frequency data window contains at least 5 to 6 cycles of period to get a resonably good estimate.

For pitch period of 120 samples, referring to Table 4.3, it has been observed in open phase that data window length 50 is comparable to open phase duration and hence estimates of the first formant has worsened there.

(ii)Formant estimation of vowels having different formant spacing:

Different synthetic vowels namely 'ae', 'u' and 'a' are constructed by setting their respective formant frequencies given in the book [12].These frequencies are also mentioned in Table 4.4. To distinguish open phase formants from closed phase we kept open phase formant frequency 60 Hz more than closed phase formant frequency. Bandwidth is kept at 100 Hz both in closed and open phase. LF model is used as voice source excitation. Table 4.4 shows the estimated formant frequency obtained after inverse filtering of simuleted vowels.

Estimation from closed phase : Formant estimation of vowel 'ae' where interformant spacing is uniformly seperated gives estimated value within 2 Hz of true value.For the vowel 'u' as the first two interformant spacing is narrower estimated second formant deviates 4 Hz from true value. For vowel 'a' as the first two formant spacing is closest estimated first and second formant deviates 4 Hz and 6 Hz respectively from true value. For all the three vowels 3rd formant is well seperated and hence estimated values are not deviated much from true value.

Estimation from open phase: Formant estimated from open phase deviates more from true value than in closed phase. A general observation is that as frequency to be estimated is lower estimated frequency is more deviated for a particular window length. If we compare the first formant estimation amongst three different vowels 'ae', 'u' and 'a' we see that for vowel 'u' deviation is largest being lowest in frequency.

(iii)Root mean square error in frequency to estimate formants from signal with additive white gaussian noise:

Here we have estimated r.m.s frequency error due to additive white gaussian noise at different signal to noise ratio. Measurment is done over 60 samples from three

consecutive pitch periods having 20 samples from each pitch period in closed phase. Formant energy decreases as one go towards higher formant and hence 3rd formant is most affected by the noise. Results are shown in the Table 4.5. This result for the first formant can be compared with the result obtained for single formant case given in the Table 4.2. Both the table shows first formant frequency gives same rms error under the white gaussian noise. Open phase measurment is not done because too many formants are missing in that phase.

Fig. 4.5 shows comparison of normalized p-error waveform at different SNR. This figure shows that with lower signal to noise ratio it is difficult to extract the instants of significant excitation. But if one has previous knowledge about the location of instant of significant excitation it is possible to extract the glottal volume velocity (g-v-v) waveform from inverse filtering of the noisy speech waveform by correctly placing the window in the closed phase. The g-v-v obtained in this way is shown in the Fig. 4.6. Formant frequency tracks corresponding to SNR=20 dB is also shown. Refering to Table 4.5, it is observed that rms error of third formant has been jumped up when SNR is decreased from 50 dB to 40 dB. It is already mentioned that it is difficult to find out the instants of significant excitation at 40 dB or at lesser SNR. This lead to the conclusion that LPC method breaks down at 40 dB SNR.

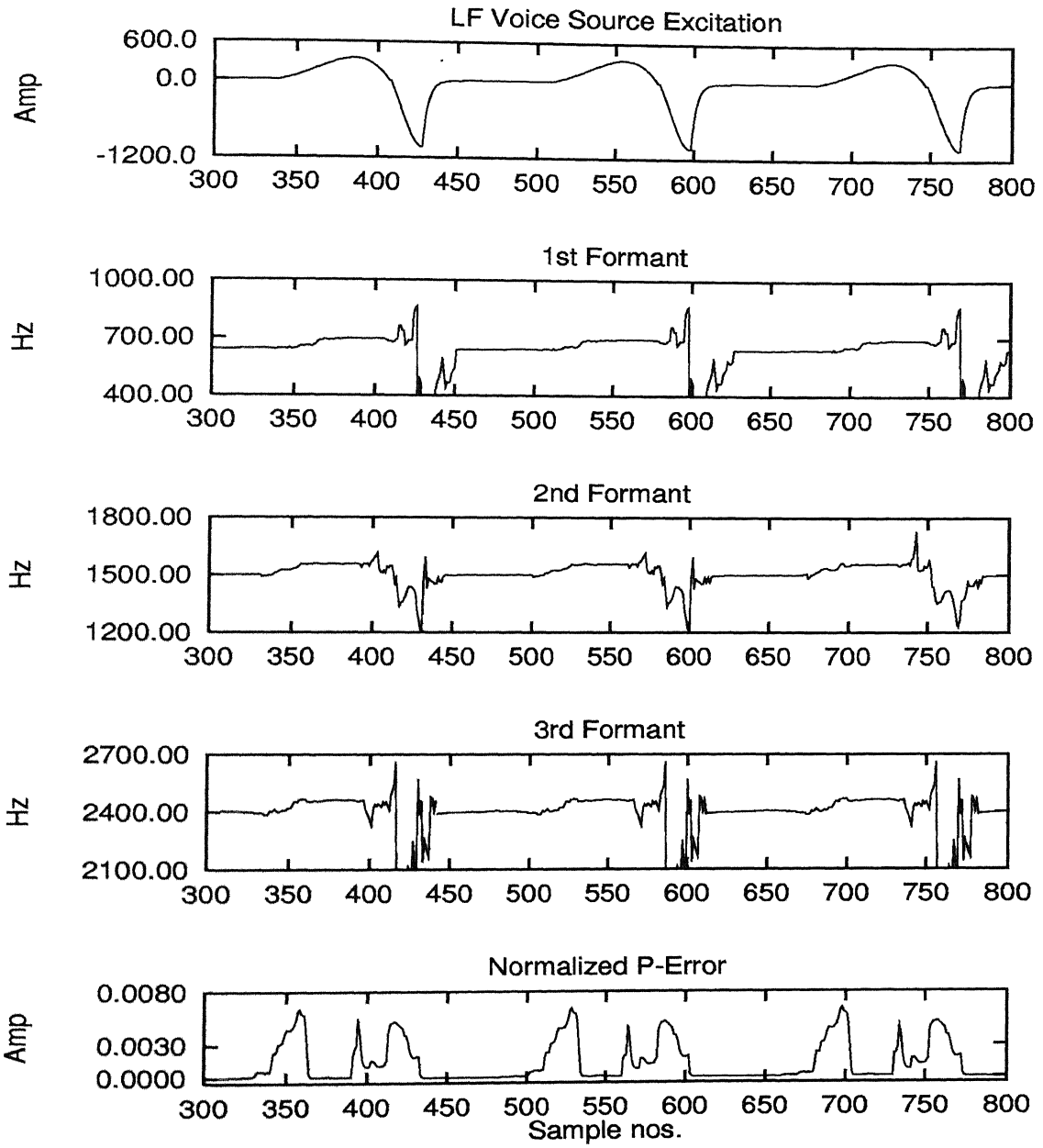


Figure 4.3: Formant frequency tracks from simulated vowel 'ae'

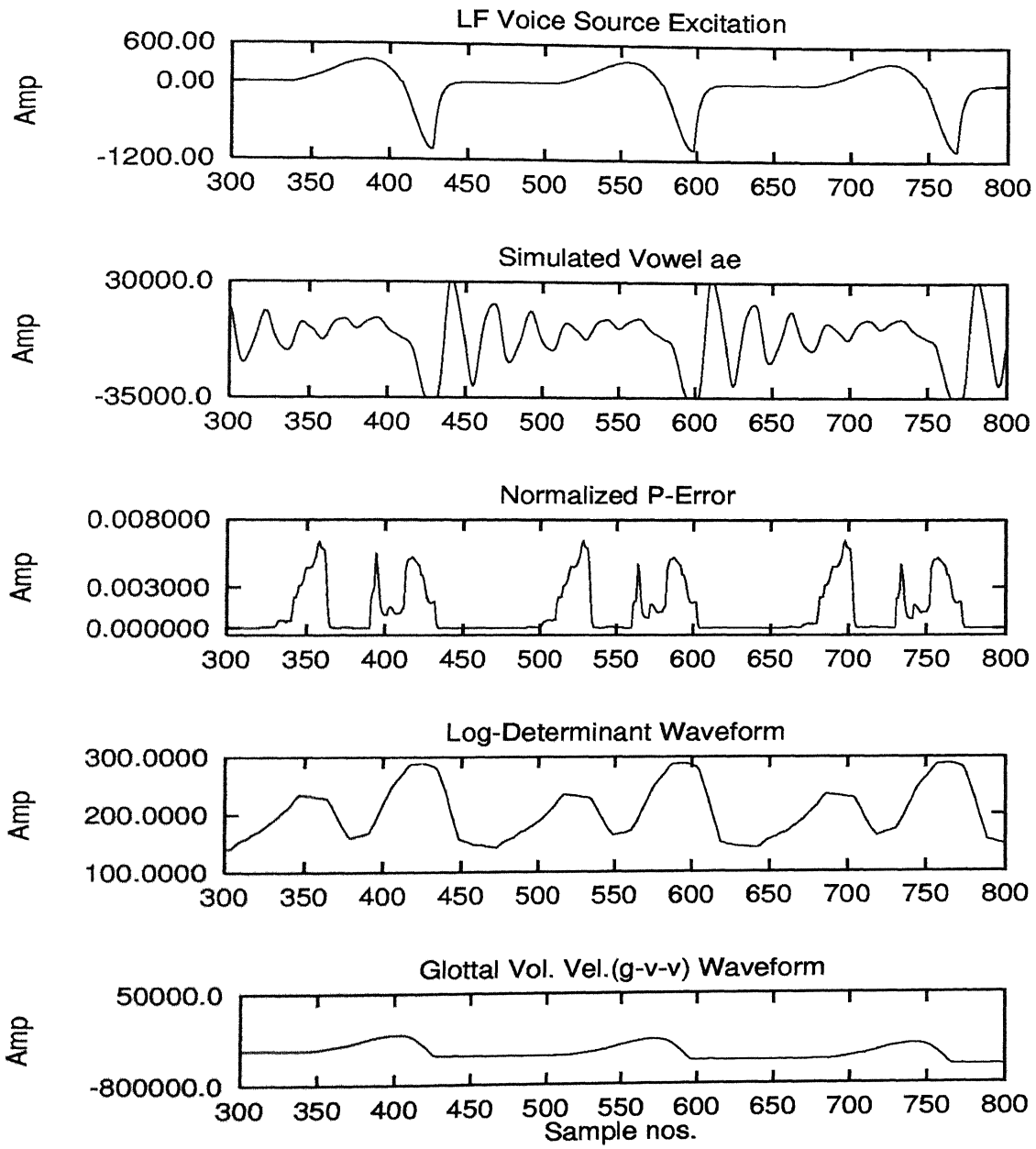


Figure 4.4: Comparisons of different waveforms from the output of inverse filter corresponding to simulated vowel 'ae'

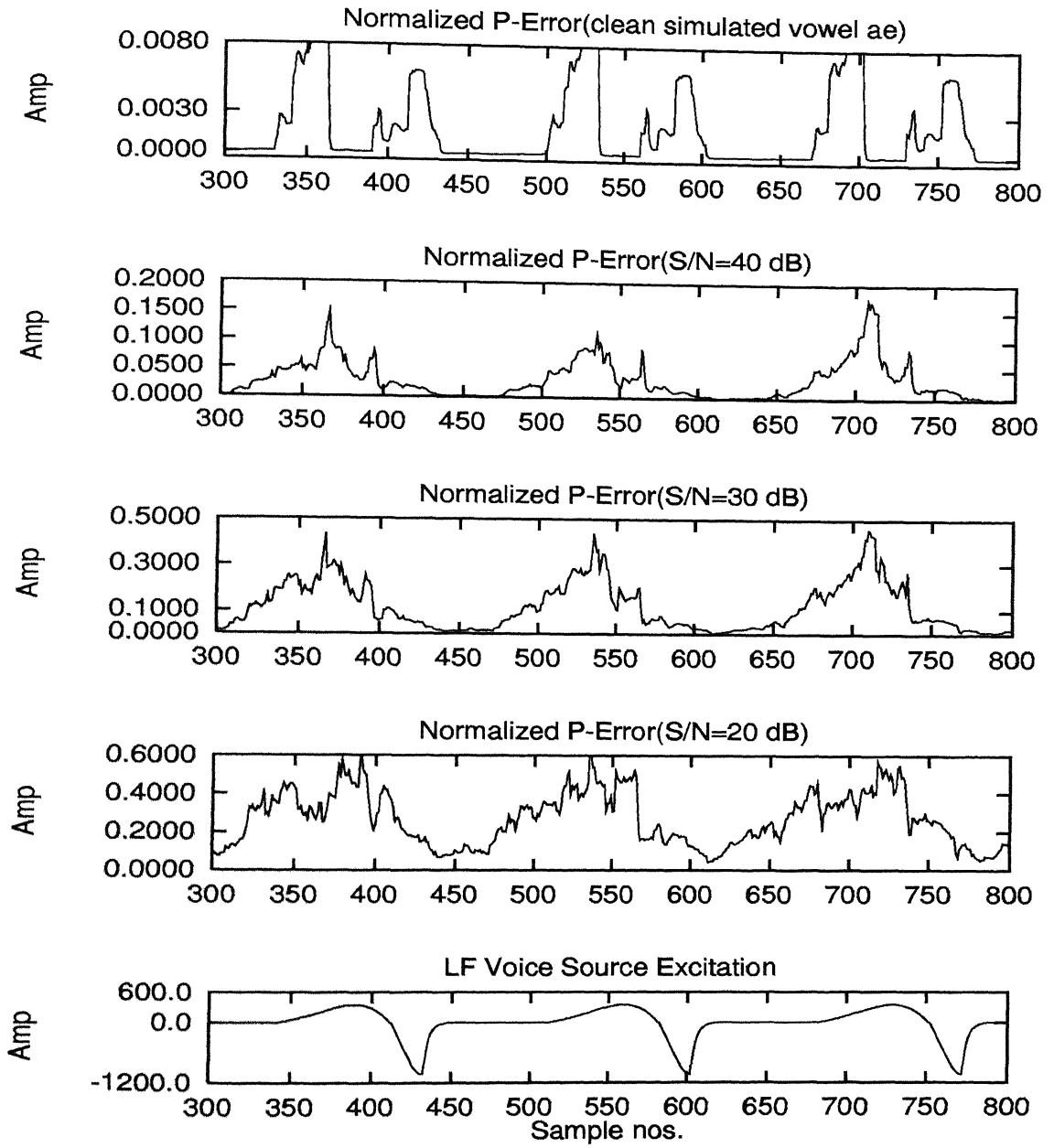


Figure 4.5: Comparison of normalized p-error waveforms at different SNR

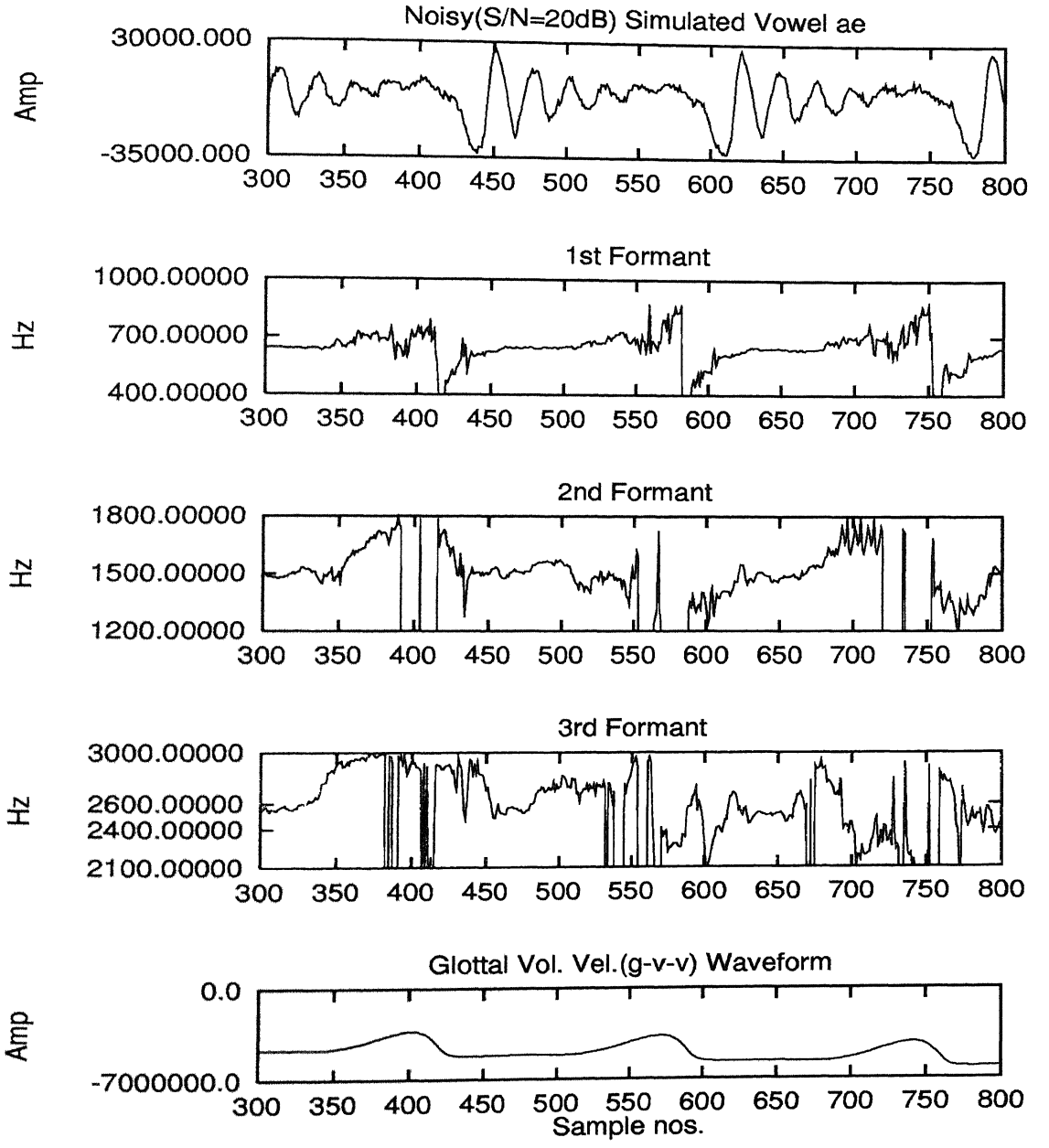


Figure 4.6: Formant frequency tracks and g-v-v from simulated vowel 'ae' with additive white Gaussian noise at S/N=20 dB

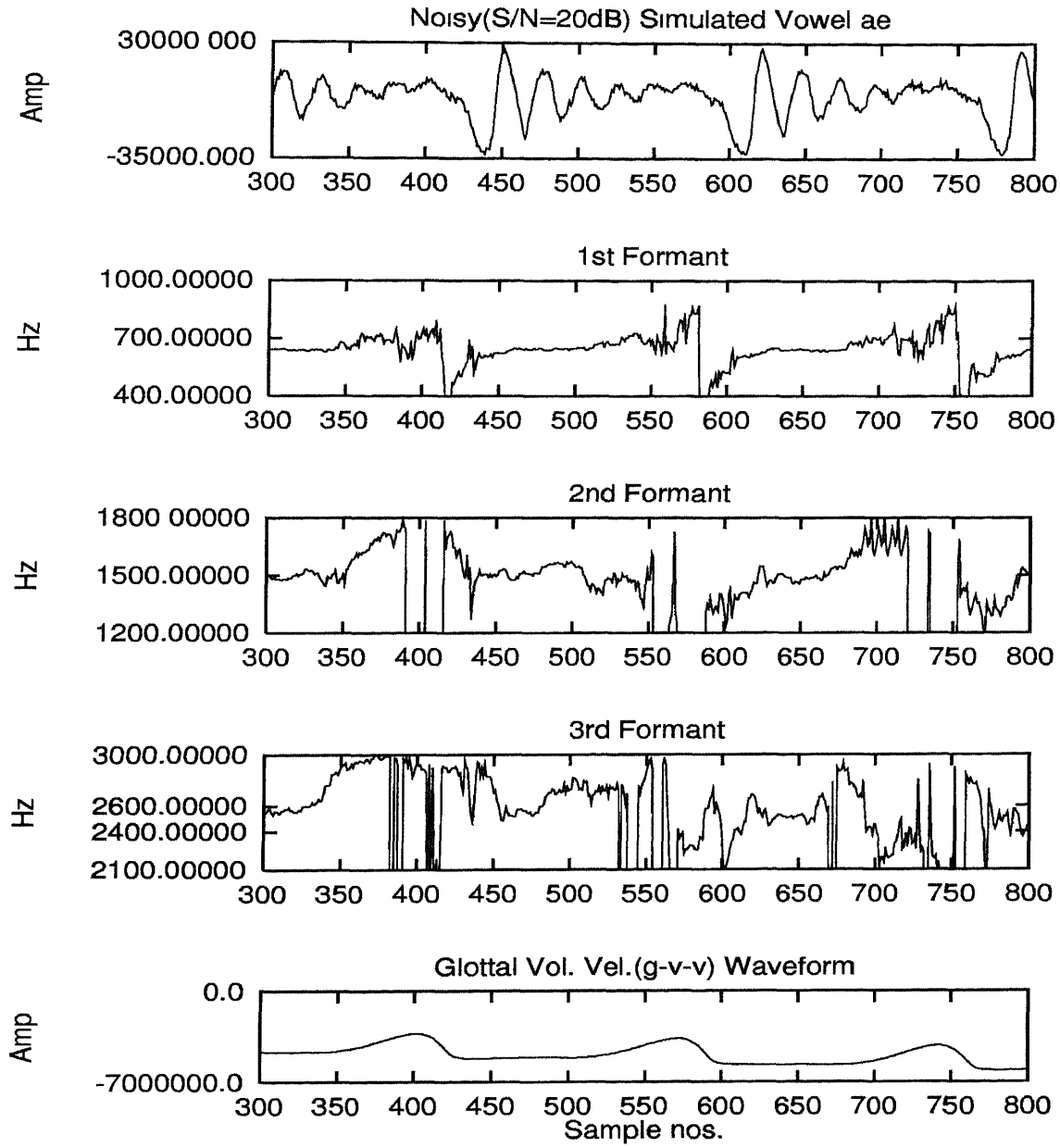


Figure 4.6· Formant frequency tracks and g-v-v from simulated vowel 'ae' with additive white Gaussian noise at S/N=20 dB

Table 4.3· Formants estimation of vowel 'ae' having different pitch periods with various data window lengths

(True formants in closed phase are f_1=640Hz, f_2=1500Hz, f_3=2400Hz; True formants in o.p. = True formants in c p. + 60 Hz ; Bandwidth is 100 Hz both in o.p and c.p.; Predictor Order = 18; Data window size = 50)

Pitch Period in samp.	Data Window	Formant	Estimated formant frequency in Hz					
			Samp.loc. c.p.			Samp.loc. o.p.		
			454 $m_{cp}-16$	470 m_{cp}	486 $m_{cp}+16$	528 $m_{op}-16$	544 m_{op}	560 $m_{op}+16$
170	50	f_1	641	641	643	689	693	695
		f_2	1500	1500	1502	1553	1559	1559
		f_3	2398	2400	2402	2469	2461	2455
	40	f_1	641	641	643	691	693	693
		f_2	1500	1500	1498	1557	1559	1559
		f_3	2398	2400	2402	2467	2459	2463
120	50	f_1	641	641	643	688	627	664
		f_2	1500	1500	1506	1552	1522	1508
		f_3	2398	2400	2414	2451	2449	2471
	40	f_1	641	641	643	691	693	709
		f_2	1500	1500	1500	1553	1551	1547
		f_3	2398	2398	2400	2467	2455	2457

Table 4 4: Formant estimation of vowels with different formant spacing

(True formants in o.p. = True formants in c.p. + 60 Hz; Bandwidth is 100 Hz both in o.p and c.p.; Predictor Order = 18; Data window size = 50; Pitch period = 170)

Vowel	True Formants in c.p	Estimated formant frequency in Hz					
		Samp.loc.at c.p.			Samp.loc.at o.p.		
		454 $m_{cp}-16$	470 m_{cp}	486 $m_{cp}+16$	528 $m_{op}-16$	544 m_{op}	560 $m_{op}+16$
ae	640	641	641	643	689	693	695
	1500	1500	1500	1502	1553	1559	1559
	2400	2398	2400	2402	2469	2461	2455
u	440	437	439	441	478	475	482
	1020	1016	1016	1018	1068	1076	1066
	2240	2240	2240	2242	2309	2297	2297
a	730	734	736	740	791	789	787
	1090	1082	1084	1082	1141	1139	1135
	2440	2443	2441	2434	2498	2498	2492

Table 4.5· Root mean square error in frequency to estimate formants from three formant signal with additive white gaussian noise
(True formants in closed phase are f_1=640Hz, f_2=1500Hz, f_3=2400Hz ;
Predictor Order = 18; Data window size = 50)

Voice Source Excitation	S/N Power	Formant	r.m.s error in closed phase (σ_{cp})
LF	50 dB	f_1	1.24
		f_2	1.46
		f_3	3.02
	40 dB	f_1	2.02
		f_2	4.63
		f_3	28.27
	30 dB	f_1	3.17
		f_2	9.89
		f_3	72.60
	20 dB	f_1	6.00
		f_2	20 74
		f_3	138 38

4.6 Conclusions

To extract the dynamical characteristics of the vocal tract system accurate analysis of speech signal is essential. Such an analysis is possible by proper choice of size and position of the analysis frame. Synchronising analysis frame with the instant of

glottal closure gives highly consistant estimates of formant frequencies Post excitation (c.p) mesurment is more accurate than pre excitation (o p.) measurment. It is also desirable to choose the frame size such that the analysis frame fits in the open or closed glottal phases. Pre excitation formant frequency tracks are less consistent than post excitation formant frequency tracks and more formants are missed in pre excitation formant tracks.

Stability of the inverse filter $H(z)$ is checked by checking the roots of the denominator polynomial $A(z)$ which all lies inside the unit circle.

The effect of ripple could not be simulated as observed in the paper Childer and Wong [1], by increasing the first and second formant bandwidths of the vocal tract by a factor of four for the open interval over that for closed interval.

Also simulation does not include the contribution due to the effects of yielding vocal tract side walls, a major contributor to broadening of the lower order formant bandwidths, the effect of viscous and thermal lossess which tends to lower the higher formant frequency and the effect of longitudinal motion of the vocal fold due to which air leaks through glottis cannot be ruled out in natural speech.

Chapter 5

Experimental Results on Natural Speech

5.1 Analysis Method

LPC covariance method is applied on natural vowels obtained from TIMIT speech database. Procedure in finding out formants from open and closed phase based on the analysis developed in last two sections. To see the variability of formants under same speaker having a specific dialect region we have chosen test/dr2/mrgg0 directory from which vowels have been picked up from different sentences type such as 'sa', 'sx' and 'si'. The results are given in the following order:

Sample 1 : dr2/mrgg0/sa1/ae(7060-9357)

Sample 2 : dr2/mrgg0/sa2/ae(5640-8520)

Sample 3 : dr2/mrgg0/sx119/ae(20706-22040)

Sample 4 : dr2/mrgg0/sa2/iy(64198-66520)

Sample 5 : dr2/mrgg0/sx299/iy(42280-44490)

Sample 6 : dr2/mrgg0/si1829/iy(64198-66520)

Sample 7 : dr2/mrgg0/si1199/ow(7060-10040)

Sample 8 : dr2/mrgg0/si11992/ey(24922-28120)

Sample 9 : dr2/mrgg0/si1829/aa(6680-9320)

Above vowels are pronounced as in the words given below.
'ae' as in 'bat' ; 'iy' as in 'beet' ; 'ow' as in 'bought' ; 'ey' as in 'buy' ; 'aa' as in 'bob' ;

Natural vowel speech signals given above are inverse filtered to find prediction coefficients and with the help of those prediction coefficients linear prediction spectrum is formed. Then by peak picking of LP spectrum first three formants are obtained. Continuously varying formant trajectories are obtained from a sliding short data window sample by sample. In addition to determine pitch period from Log-determinant waveform closed and open phase duration are also estimated which in turn helps to determine the size of data window. Data window size chosen in this way is 50. Since sampling frequency is 16 KHz predictor order chosen is 18. Formant frequency measurments in open and closed phase are taken in three consecutive pitch periods corresponding to locations where prediction error is mimimum. These minimum error locations are also the positions where window completely fits within the closed and open phases. m_{cp} and m_{op} represents these positions in the Tables given later.

5.2 Observations and Conclusions

Formant frequencies obtained from open phase and closed phase are also respectively called pre and post excitation formant frequency. Results corresponding to natural vowels are shown in the Tables follow. It has been observed from natural speech for certain vowels that pre and post-excitation formant frequencies can differ significantly and sometimes not. In general most consistent results are obtained from post-excitation analysis frames because glottal airflow is smallest or zero immediately after glottal closure. Non-linear coupling between sub and the supra-glottal tract are nil and their influence on the formant frequencies are not present.

In several cases of natural speech vowels, clear increases of formant frequency are observed in the open phase compared with the closed phase.

Vowels having closely spaced formants,for example vowel 'a', apparent merging of f_1 and f_2 in open phase is observed due to influence of the excitation, though in close phase they are well separated due to the absence of excitation. This is given in Table 5.9.

It has also been observed that formant frequency varies from pitch period to period. Since we have shifted window sequentially sample by sample frame rate is quite high and more data per unit time are obtained so that occasional omissions of formants is minimised and fast transitional behaviour of formants, if any, can be observed.

Table 5.1: Formant estimation of natural vowel 'ae' (sample1) from different pitch periods using data window size = 50 and predictor order = 18

(True formants given [12] $f_1 = 660$ Hz, $f_2 = 1720$ Hz and $f_3 = 2410$ Hz)

Vowel	pitch period	Estimated formant frequency in Hz from					
		1st Period		2nd Period		3rd Period	
	Sample number	c p	o p	c.p.	o p.	c.p.	o p.
		255	322	378	445	501	568
		m_{cp}	m_{op}	m_{cp}	m_{op}	m_{cp}	m_{op}
ae	f_1	590	592	592	590	598	566
	f_2	1695	1695	1680	1678	1678	1629
	f_3	2488	2693	2475	2662	2498	2662

Table 5.2: Formant estimation of natural vowel 'ae' (sample2) from different pitch periods using data window size = 50 and predictor order = 18

(True formants given [12] $f_1 = 660$ Hz, $f_2 = 1720$ Hz and $f_3 = 2410$ Hz)

Vowel	pitch period	Estimated formant frequency in Hz from					
		1st Period		2nd Period		3rd Period	
	Sample number	c.p.	o p.	c.p.	o.p.	c.p.	o.p.
		150	211	290	355	425	495
		m_{cp}	m_{op}	m_{cp}	m_{op}	m_{cp}	m_{op}
ae	f_1	535	643	553	643	570	666
	f_2	1672	1646	1658	1639	1652	1619
	f_3	2377	2408	2371	2408	2377	2316

Table 5.3: Formant estimation of natural vowel 'ae' (sample3) from different pitch periods using data window size = 50 and predictor order = 18

(True formants given [12] $f_1 = 660$ Hz, $f_2 = 1720$ Hz and $f_3 = 2410$ Hz)

Vowel	pitch period	Estimated formant frequency in Hz from					
		1st Period		2nd Period		3rd Period	
	Sample number	c.p	o.p.	c.p	o p.	c.p.	o.p.
		302	372	456	526	610	680
		m_{cp}	m_{op}	m_{cp}	m_{op}	m_{cp}	m_{op}
ae	f_1	617	775	639	826	652	789
	f_2	1393	1467	1410	1463	1420	1480
	f_3	1963	2074	2021	2035	2045	2055

Table 5.4: Formant estimation of natural vowel 'iy' (sample4) from different pitch periods using data window size = 50 and predictor order= 18

(True formants given [12] $f_1 = 270$ Hz, $f_2 = 2290$ Hz and $f_3 = 3010$ Hz)

Vowel	pitch period	Estimated formant frequency in Hz from					
		1st Period		2nd Period		3rd Period	
	Sample number	c.p.	o.p.	c.p.	o.p.	c.p.	o.p.
		140	211	291	362	442	513
		m_{cp}	m_{op}	m_{cp}	m_{op}	m_{cp}	m_{op}
iy	f_1	436	498	422	488	412	514
	f_2	1803	1900	1895	2012	2010	2164
	f_3	3342	3340	3316	3348	3350	3412

Table 5 5: Formant estimation of natural vowel 'iy' (sample5) from different pitch periods using data window size = 50 and predictor order= 18

(True formants given [12] $f_1 = 270$ Hz, $f_2 = 2290$ Hz and $f_3 = 3010$ Hz)

Vowel	pitch	Estimated formant frequency in Hz from					
	period	1st Period		2nd Period		3rd Period	
	Sample	c.p.	o.p.	c.p.	o.p.	c.p.	o.p.
	number	18	116	171	267	326	418
		m_{cp}	m_{op}	m_{cp}	m_{op}	m_{cp}	m_{op}
iy	f_1	414	439	400	461	377	473
	f_2	2299	2379	2277	2344	2248	2383
	f_3	3254	3342	3289	3287	3297	3191

Table 5 6: Formant estimation of natural vowel 'iy' (sample6) from different pitch periods using data window size = 50 and predictor order= 18

(True formants given [12] $f_1 = 270$ Hz, $f_2 = 2290$ Hz and $f_3 = 3010$ Hz)

Vowel	pitch	Estimated formant frequency in Hz from					
	period	1st Period		2nd Period		3rd Period	
	Sample	c.p.	o.p.	c.p.	o p.	c.p	o.p.
	number	142	210	283	354	424	491
		m_{cp}	m_{op}	m_{cp}	m_{op}	m_{cp}	m_{op}
iy	f_1	301	307	301	328	297	377
	f_2	2086	2143	2096	2131	2107	2176
	f_3	2918	3001	2902	2906	2887	2898

Table 5.7: Formant estimation of natural vowel 'ow' (sample7) from different pitch periods using data window size = 50 and predictor order = 18

(True formants given [12] $f_1 = 570$ Hz, $f_2 = 840$ Hz and $f_3 = 2410$ Hz)

Vowel	pitch period	Estimated formant frequency in Hz from					
		1st Period		2nd Period		3rd Period	
	Sample number	c.p.	o.p.	c p	o.p.	c.p	o p
		125	184	261	320	397	456
		m_{cp}	m_{op}	m_{cp}	m_{op}	m_{cp}	m_{op}
ow	f_1	525	531	521	520	512	523
	f_2	1043	1061	1035	1117	1029	1080
	f_3	2674	2697	2660	2713	2641	2623

Table 5.8: Formant estimation of natural vowel 'ey' (sample8) from different pitch periods using data window size = 50 and predictor order = 18

(True formants given [12] $f_1 = 530$ Hz, $f_2 = 1840$ Hz and $f_3 = 2480$ Hz)

Vowel	pitch period	Estimated formant frequency in Hz from					
		1st Period		2nd Period		3rd Period	
	Sample number	c.p.	o.p.	c.p.	o.p.	c.p.	o.p.
		160	234	326	394	492	554
		m_{cp}	m_{op}	m_{cp}	m_{op}	m_{cp}	m_{op}
ey	f_1	480	494	477	492	467	490
	f_2	1744	1826	1773	1883	1807	1922
	f_3	2484	2518	2488	2512	2496	2561

Table 5.9: Formant estimation of natural vowel 'aa' (sample9) from different pitch periods using data window size = 50 and predictor order = 18

(True formants given [12] $f_1 = 730$ Hz, $f_2 = 1090$ Hz and $f_3 = 2440$ Hz)

Vowel	pitch period	Estimated formant frequency in Hz from					
		1st Period		2nd Period		3rd Period	
	Sample number	c p.	o.p.	c p	o.p.	c.p	o.p.
		322	394	474	540	626	682
		m_{cp}	m_{op}	m_{cp}	m_{op}	m_{cp}	m_{op}
aa	f_1	678	842	682	852	688	865
	f_2	1066	-	1082	-	1096	-
	f_3	2758	2742	2721	2727	2695	2665

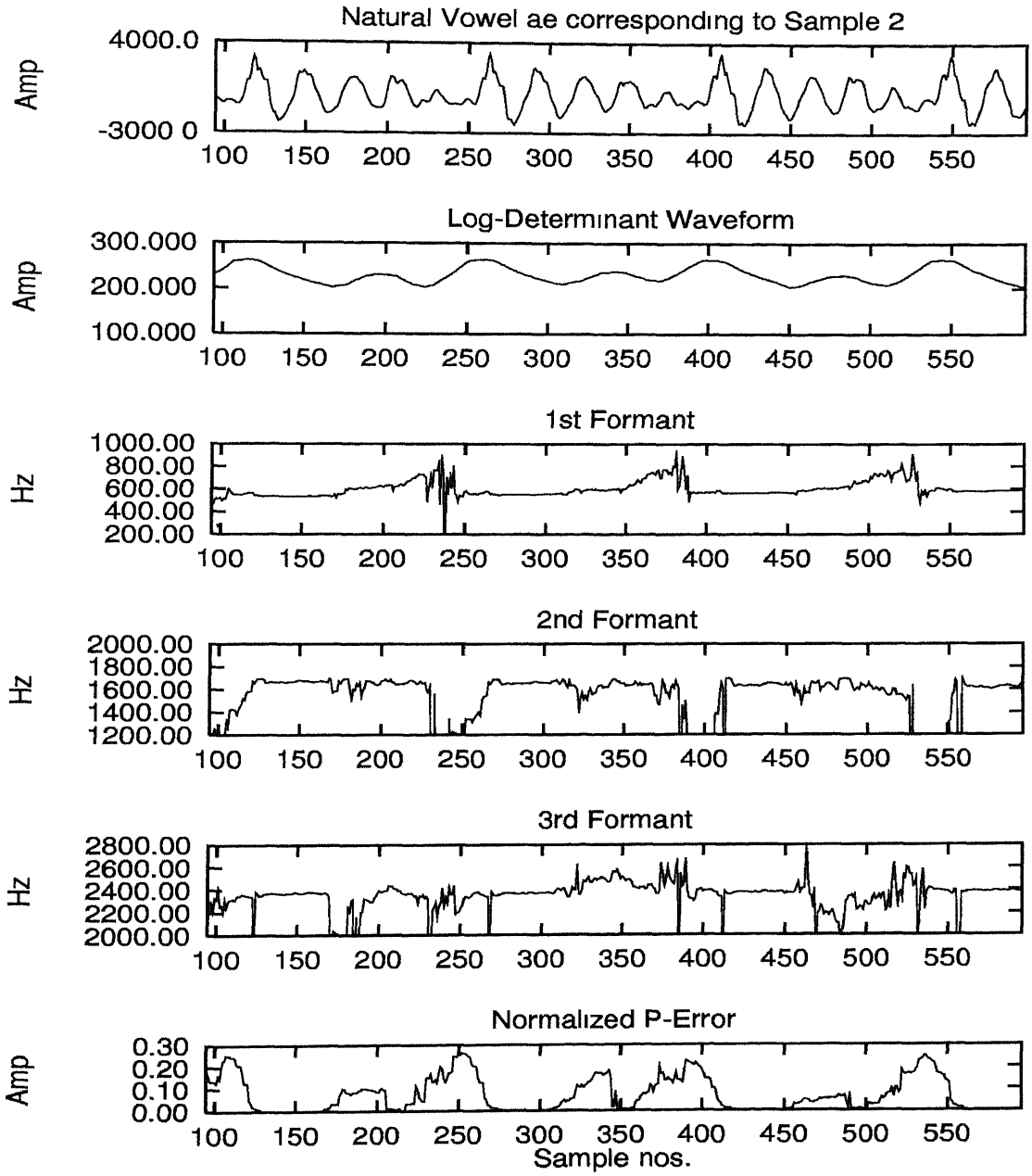


Figure 5.1: Output waveforms of inverse filter corresponding to natural vowel 'ae'

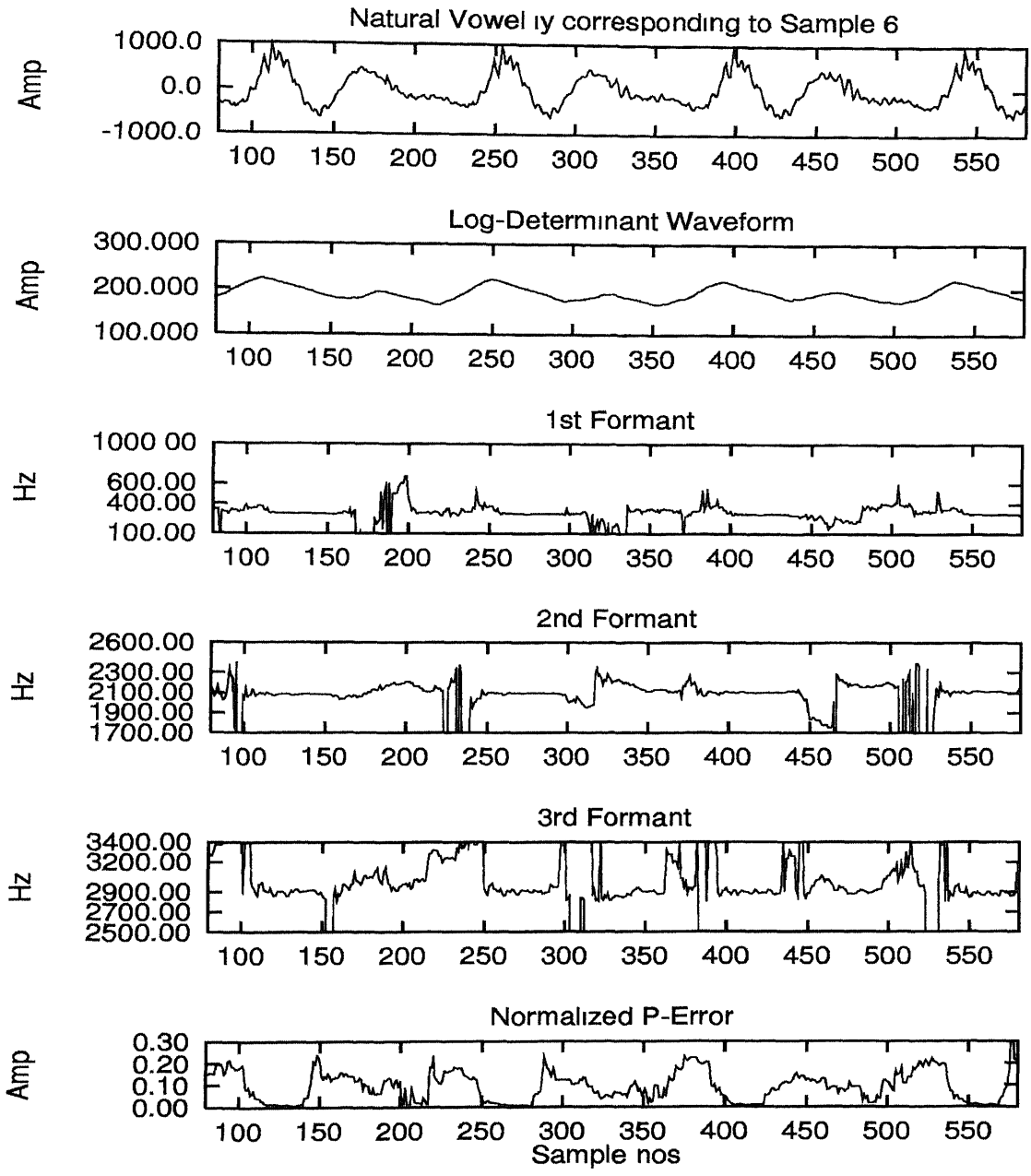


Figure 5.2: Output waveforms of inverse filter corresponding to natural vowel 'iy'

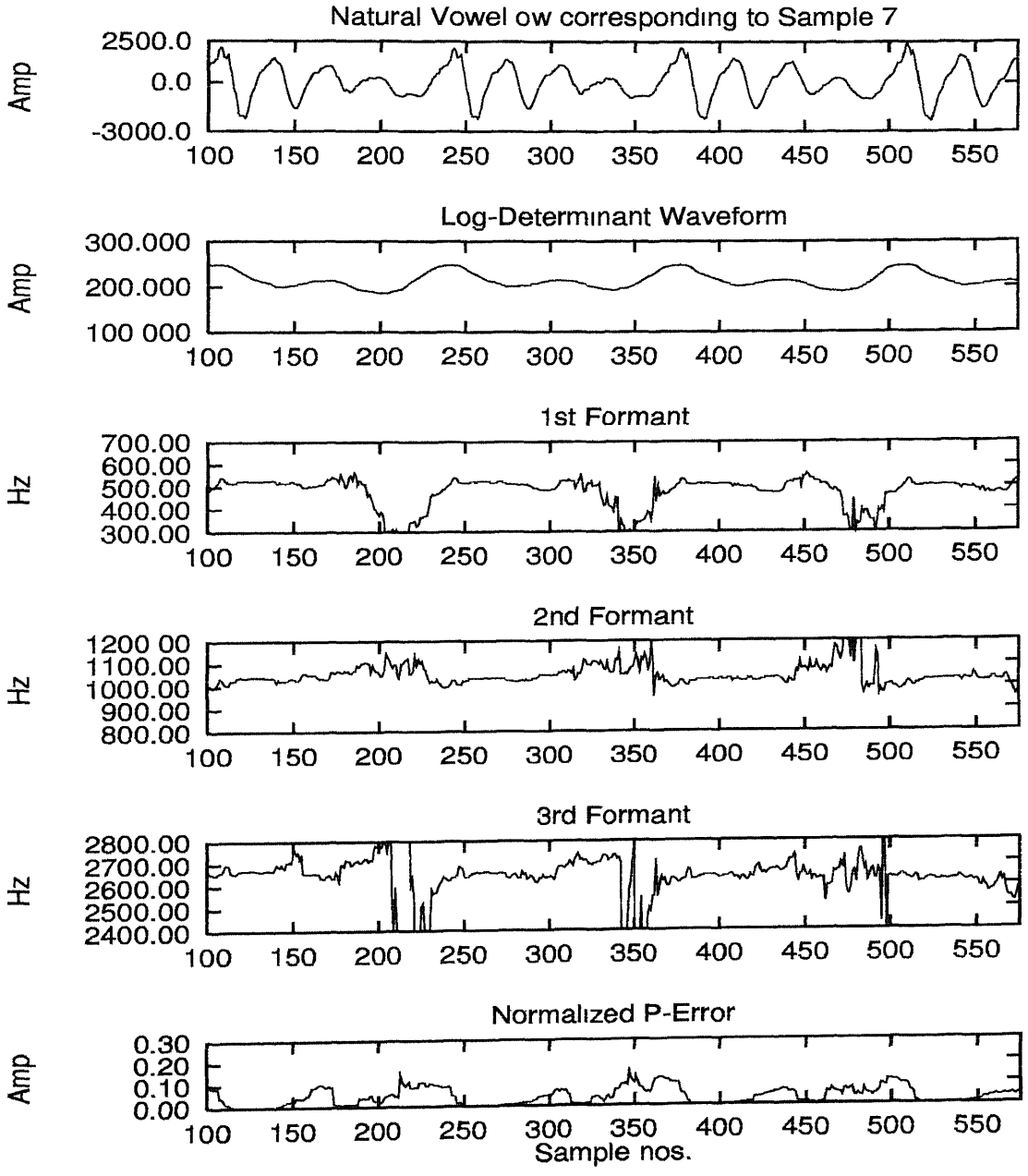


Figure 5.3: Output waveforms of inverse filter corresponding to natural vowel 'ow'

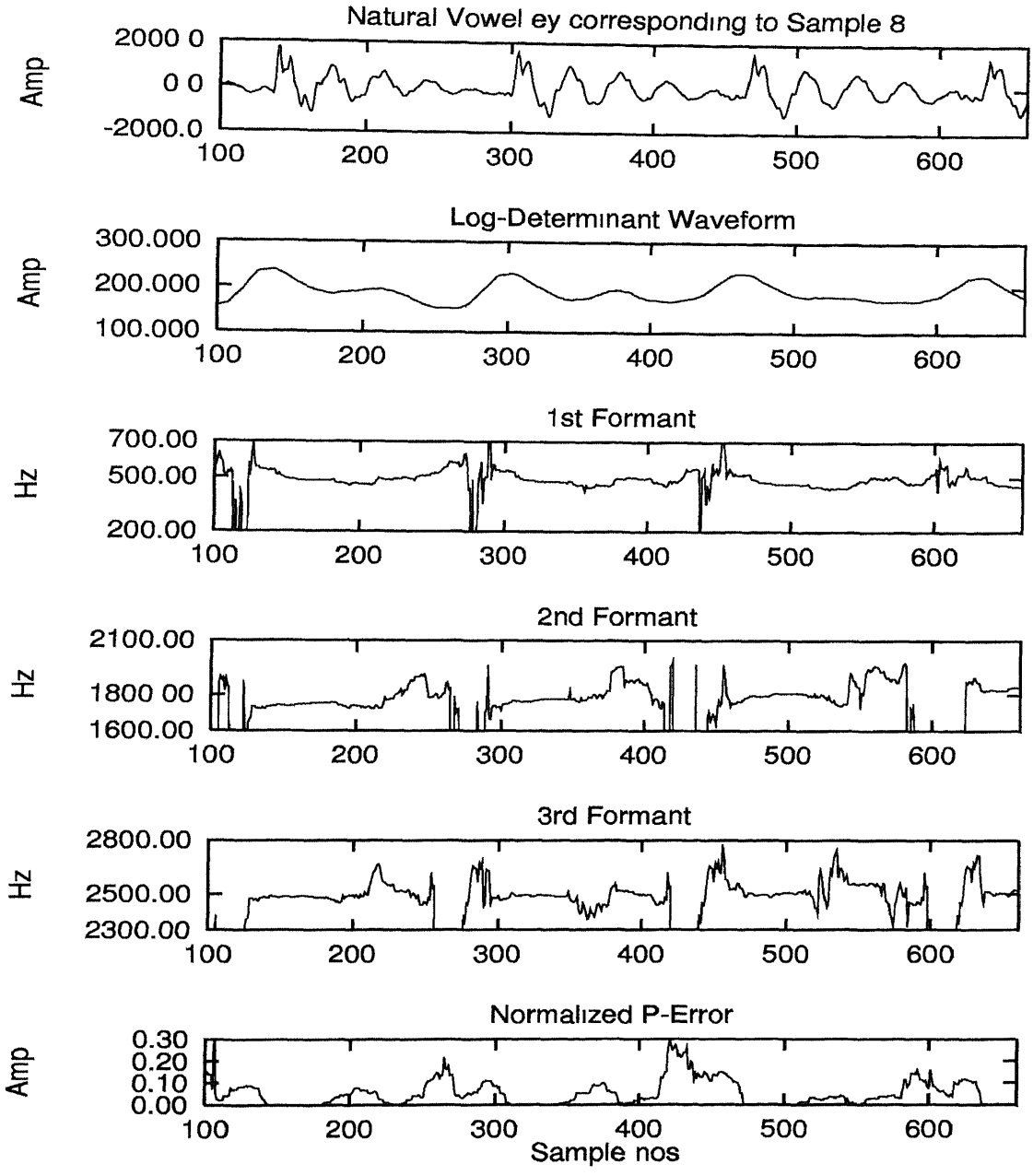


Figure 5.4: Output waveforms of inverse filter corresponding to natural vowel 'ey'

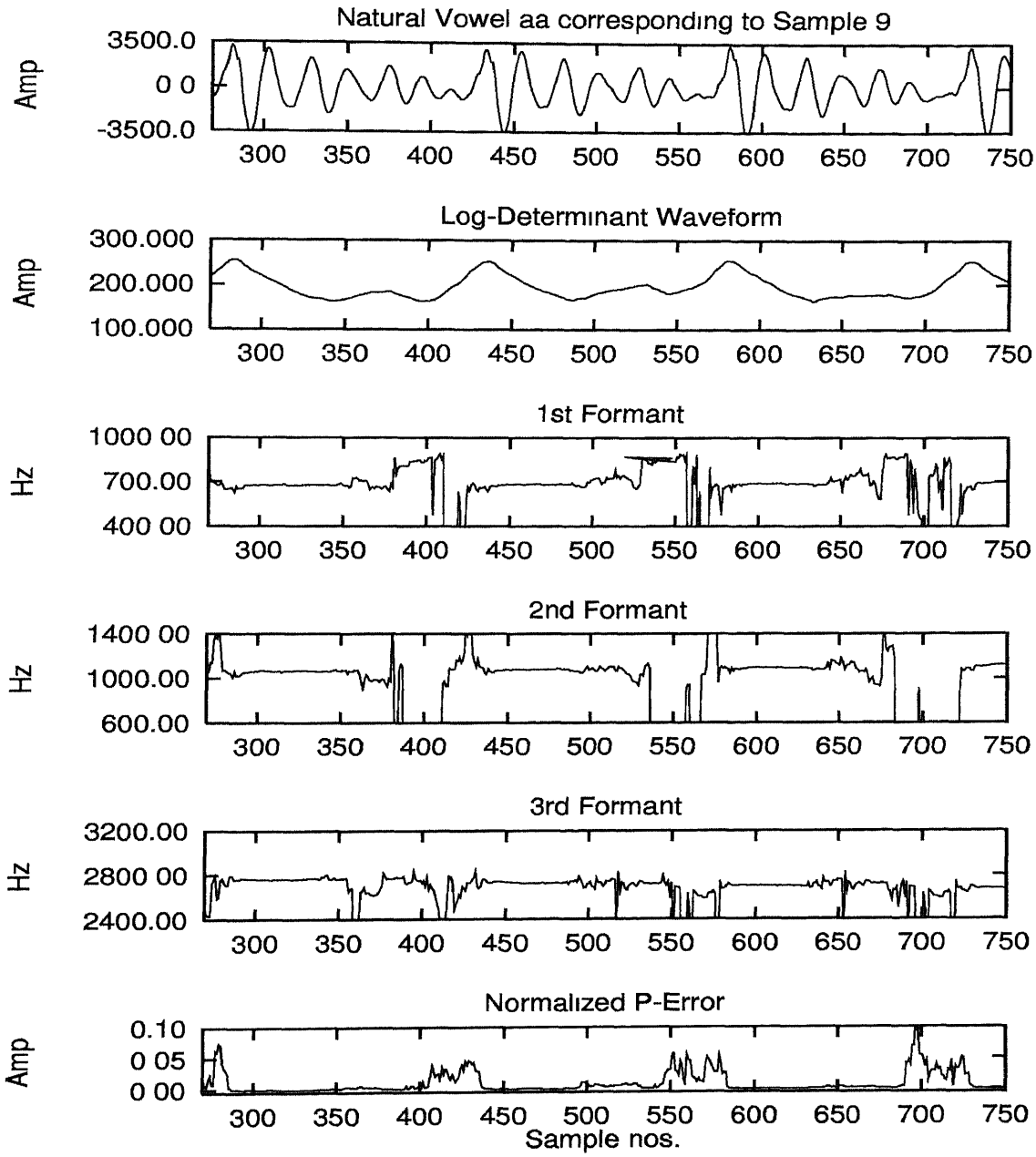


Figure 5.5: Output waveforms of inverse filter corresponding to natural vowel 'aa'

5.3 Future Work

Peak picking method in finding out formant peaks is a more tractable problem with linear prediction spectra than with other forms of spectral analysis because spurious peaks are rare. However peak picking has a limitation of merged peaks and cannot distinguish very closely spaced formants. Hence to track this problem formant enhancement technique should be incorporated in the modified algorithm.

A natural extention of this work would be to investigate formants behaviour under various source tract interaction effects such as skewing, ripples, superposition, supra-glottal effect etc. For this one has to suitably model the excitation.

Another effect to be studied is the behaviour of formant bandwidths under glottal loss as glottal source impedance has significant effects upon formant bandwidth, for example glottal source impedance decreases with decreasing formant frequency and there will be glottal loss for low frequency formants.

Bibliography

[1] Childers and Wong, "Measuring and Modelling Vocal Source Tract Interaction", *IEEE Trans on Biomedical Engineering*, Vol 41, No. 7, pp 663-671, July 1994

[2] Yegnanarayana and Veldhuis, "Vocal Tract System Characteristics", *IEEE Trans. on Speech and Audio processing*, Vol. 6, No. 41, pp 314-326, July 1998

[3] Dennis H. Klatt , "Software for a Cascade/Parallel Formant Synthesizer", *Journal of the Acoustic society of America*, Vol. 67, No. 3, pp 971-994, March 1980

[4] David Y. Wong, J. D. Markel and A. H. Gray JR., "Least Squares Glottal Inverse Filtering ", *IEEE Transanctions on ASSP*, Vol. 27, No. 4, pp 350-355, Aug. 1979

[5] S.Parthasarathy and D. W. Tufts, "Synchronous Modelling of Voiced Speech", *IEEE Transanctions on ASSP*, Vol. 35, No. 9, pp 1241-1249, Sept. 1987

[6] John D. Markel, "Digital Inverse Filtering", *IEEE Trans.on Audio and Electroaccoustics*, Vol. 20, No. 2, pp 129-137, June 1972

[7] H. W. Strube, "Determination of the Instants of Glottal Closure ", *Journal of the Acoustic society of America*, Vol. 56, No. 5, pp 1625-1629, Nov 1974

[8] Gray and Markel, "Speech Analysis",
IEEE Transanctions on ASSP, Vol. 21, No. 2, pp 214-216, June 1974

[9] Stephanie S. Mccandless ,
"An Algorithm for Automatic Formant Extraction Using LP Spectra "
IEEE Transanctions on ASSP, Vol. 22, No. 2, pp 214-216, April 1974

[10] J. Makhoul, "Linear Prediction : A Tutorial Review",
Proceedings of the IEEE, Vol. 63, No. 4, pp 561-580, April 1975

[11] S Taniguchi, "Glottal Source Tract Interaction",
Journal of the Acoustic society of America, Vol. 78, No. 5, pp 1541-1550, 1985

[12] L. R. Rabiner/R. W. Schafer , "*Digital Processing of Speech Signal*",
Printice-Hall Inc., 1978

[13] William H.Press and Saul A.Tuolosky and William T Vellerling and Brain P. Flannery, "*NUMERICAL RECIPES in C, The Art of Scientific Computing*",
Cambridge University Press, 1996